Title
Rasch and Rationality: Scale typologies as applied to Item Response Theory

Author
Freund, Rebecca

Publication Date
2019

Peer reviewed|Thesis/dissertation

Rasch and Rationality: Scale typologies as applied to Item Response Theory

by

Rebecca Lynn Freund

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark R. Wilson, Chair
Professor Sophia Rabe-Hesketh
Professor Alan Hubbard

Summer 2019

Rasch and Rationality: Scale typologies as applied to Item Response Theory

Abstract

Rasch and Rationality: Scale typologies as applied to Item Response Theory

by

Rebecca Lynn Freund

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark R. Wilson, Chair

This dissertation consists of three papers on a central topic: the application of scale type theory to the Rasch and 2PL models. Each paper uses a different framework or set of frameworks for defining a typology of scales.

In the first paper, I begin the analysis of scale types through the Stevens (1946) typology. I introduce the notion of difference scales using a formalization of this typology. In applying this formalization to the Rasch and 2PL models, I discuss alternate paradigms for conceptualizing the restrictive assumptions of the Rasch model.

In the second paper, I apply the Suppes (1958) typology to the two IRT models. The conclusions here echo and reinforce those of the first chapter. In the second half of this paper, I examine other perspectives on connecting scale type theory and the Rasch model, focusing on the case of additive conjoint measurement theory.

The first two papers consider the scale types of the odds and logit forms of the models separately. In the third paper, I look at typologies in which the model form is not a factor. The first of these is the axiomatic system of Hölder (1901), which identifies two types of quantities: magnitudes, and points on a line. The second is the homogeneity-uniqueness typology of Narens and Luce (1986), which classifies scales by the types of choices that are possible during the process of numeral assignment.

Together, the three papers form an argument for considering psychometric properties behaving as predicted by the Rasch model as having the scale type of ratio scale attributes, while those that behave as predicted by the 2PL model have the scale type of interval scale attributes.

Contents

List of Figures

List of Tables

Acknowledgments

Thank you first to my advisor and committee chair, Prof. Mark Wilson, and to my committee members Prof. Sophia Rabe-Hesketh and Prof. Alan Hubbard, for supporting me through this process, and for being flexible as the topic evolved. Thanks also to David Torres Irribarra, and the rest of the Berkeley QME/SRM/POME community, for your insights, assistance, and friendship. Thank you to my parents and sister for always being there for me. And to my partner, Antoine Duquerrois, thank you for your encouragement, patience, and love.

Chapter 1

Applying a formalization of the Stevens typology to Item Response Theory models

1.1 Introduction

Item Response Theory models attempt to describe a respondent's probability of success on an item as a function of the respondent's proficiency level and the item's difficulty level. The Rasch model gives the log odds (logit) of success on an item as:

$$\text{logit}(x_i = 1|\theta) = \theta - \beta_i \tag{1.1}$$

where θ represents a respondent's ability and β_i represents an item's difficulty (Rasch, 1960/ 1980).

Equivalently, the Rasch model can be expressed using odds instead of log odds:

$$\text{Odds}(x_i = 1|t) = \frac{t}{b_i} \tag{1.2}$$

where t and b_i represent the θ and β_i parameters respectively, adjusted for the odds scale. If t and b_i are set to be the exponentiations of θ and β_i respectively, the two models make equivalent predictions.

The related Two Parameter Logistic (2PL) model adds a so-called item discrimination parameter α_i (Birnbaum, 1968):

$$\text{logit}(x_i = 1|\theta) = \alpha_i(\theta - \beta_i) \tag{1.3}$$

The odds formulation of the 2PL model is:

$$\text{Odds}(x_i = 1|t) = \left(\frac{t}{b_i}\right)^{\alpha_i} \tag{1.4}$$

If the 2PL model is modified such that all the discrimination parameters are constrained to be equal to a constant α_0 across items, the result is a generalized Rasch model (GRM):

$$\text{logit}(x_i = 1|\theta) = \alpha_0(\theta - \beta_i) \tag{1.5}$$

This equation is equivalent to Equation 1.1, except that the standard deviation of θ can vary. Equation 1.5 has also been referred to as describing a "family of Rasch models" (Fischer, 1995), in which each choice of discrimination parameter is considered a separate Rasch model. When the value of α_0 parameter is chosen *a priori*, it is referred to as an index rather than a parameter, and the model is a One-Parameter Logistic model (Verhelst & Glas, 1995).

These models make certain quantitative assumptions and predictions regarding how respondents will perform on items, and estimates obtained by fitting these models to data will have certain mathematical properties. The act of measurement using these models involves assigning numeric values to respondents' levels of an attribute.

Scale type theory, as originally outlined by Stevens (1946), and further formalized by Suppes and Zinnes (1963), categorizes these types of measurement assignments into different *scales*, each with different empirical and mathematical properties. The two most commonly used quantitative scales are the *interval scale* and the *ratio scale*. The goal of this paper is to apply that categorization to the IRT models through an extended formalization of the Stevens system of isomorphic determinations. The main argument of this paper is that attributes that perform as predicted by the 2PL model should be classified as an interval scale, while those that follow the Rasch model should be classified as a ratio scale (given a specific belief regarding the nature of the common slope in the Rasch model).

These categorizations depend on using the logit and odds forms, respectively, of the 2PL and Rasch. This paper also describes two additional scales, the *absolute difference* and *relative ratio* scales, which describe the alternate model forms (logit for Rasch, and odds for 2PL, respectively), and discusses the relationships between the four scale types.

1.2 The Stevens typology

1.2.1 Background

The modern concept of scale types originated in a fundamental paper by S. S. Stevens (1946). To understand Stevens' perspective on scales, and why they were so important in his work, it is first necessary to understand his views on measurement as a whole. These views were nuanced, and at times seemingly contradictory. On the one hand, the definition of measurement that he adopted from Campbell and Jeffreys (1938) and which he made famous, that of "the assignment of numerals to objects or events according to rules" (Stevens, 1946), suggests an almost limitless view of measurement. Michell (1999) complains that "[t]hose who accept Stevens' definition will believe that they can measure whenever they have a rule for assigning numerals to objects or events, regardless of whether the relevant attribute is

quantitative" and notes that "[p]rocedures for assigning numbers or numerals to objects or events according to some rule can be devised on request, and without limit." Stevens' own writings on his famous definition often reinforce this liberal view, stating explicitly that "The only rule not allowed would be random assignment" (1976, p. 47) or "provided a consistent rule is followed, some form of measurement is achieved" (1959, p. 19).

Yet, Stevens consistently paired these operational dicta with more representational definitions of measurement, writing that "measurement occurs whenever an element from one domain is matched, equated, or conjoined to an element of another domain" (1976, p. 46) and "measurement is the process of mapping empirical properties or relations into a formal model" (1959, p. 20). These definitions are narrower than the standard definition, in that they require matching or mapping operations rather than just assignment, but they also raise the possibility of a measurement process that does not involve numerical assignment or numbers at all. Stevens suggests, as a non-numeric example, asking a subject to "squeeze a hand dynamometer to signify by the force of his handgrip the apparent intensity of a light" (1976, p. 46).

Michell (1986) interprets these varying definitions as Stevens' attempt to "weld together two measurement traditions: representationalism and operationalism." Stevens reconciled these two definitions by claiming that the "according to a rule" clause was sufficient to establish the representational framework:

> Although the definition of measurement could, if we wished, be broadened to include the determination of any kind of relation between properties of objects or events, it seems reasonable, for the present, to restrict its meaning to those relations for which one or another property of the real number system might serve as a useful model. *This restriction is implied when we say that measurement is the assignment of numerals to aspects of objects or events according to rule.*(Stevens, 1959, p. 24, emphasis added)

This was how Stevens attempted to establish his definition of measurement as open enough to allow any number of practices in psychological and psychophysical measurement, but focused enough to exclude practices that resulted in measurements that were not *meaningful*. He declared that his definition was not all-permitting after all; it in fact contained within it an implication that the assignment of numerals to objects should be made such that a property of the real number system should serve as a useful model for relations between objects or events. The determination of the nature of these properties, models, and relations would serve as the basis for his work on scale types.

"Meaning" can be seen as at the heart of Stevens' work on scale types. Measurement was only meaningful when numerical relationships between the assigned numbers represented empirical relations between objects, and scale transformations should preserve these relations. More controversially for his time, Stevens held that calculated statistics were only meaningful if they maintained invariance under these transformations. Disallowed

statistics—for example, a ratio of data points on an interval scale, which will vary under linear transformations—he referred to as "meaningless."

The connection between invariance and meaningfulness was made explicit in Suppes (1958), who expanded on Stevens' ideas using the following definition:

> An empirical hypothesis or any statement in fact, which uses numerical quantities, is empirically meaningful only if its truth-value is invariant under the appropriate transformations of the numerical quantities involved.

1.2.2 The Stevens scale types

Stevens (1946) distinguished four types of scales: nominal, ordinal, interval, and ratio. For Stevens, scale type is determined by the presence of *empirical operations*. More specifically, he writes

> In dealing with the aspects of objects we invoke empirical operations for determining equality (classifying), for rank-ordering, and for determining when differences and when ratios between the aspects of objects are equal. The conventional series of numerals yields to analogous operations... The isomorphism between these properties of the numeral series and certain empirical operations which we perform with objects permits the use of the series as a *model* to represent aspects of the empirical world.
>
> The type of scale achieved depends upon the character of the basic empirical operations performed. (p. 677)

Thus, we have two types of operations: Empirical operations, acting on the properties of the objects (represented by $A, B, C, D \in \boldsymbol{X}$), and numeric operations, acting isomorphically on the numerals assigned to these aspects (represented by $f(A), f(B), f(C), f(D) \in \mathbb{R}$, where f is a map from $\boldsymbol{X}$ to $\mathbb{R}$). In a psychometric context, the empirical operations would directly involve respondents' construct ability, tendency, or attitude, while the numeric operations would be applied to the numerals assigned to their locations on the scale.

Stevens' empirical operations are all types of determinations: Determination of equality, determination of lesser or greater, determination of equal ratios or differences. The numeric operations then represent mathematical analogues of these real-world relations. Scale type selection is prompted by the available empirical operations, and achieved when there is isomorphism between the empirical relations and numeric operations (i.e., when the numeric relations on $f(A), f(B), f(C), f(D)$ hold if and only if the respective empirical determinations on A, B, C, D hold).

For each scale type, Stevens outlines the numerical transformations that are permissible on the numeral series. These transformations are sufficient to define the nature of the scale. But for Stevens, the idea that some transformations are "allowable" is only sensical if they preserve an empirical relationship:

> The permissible transformations defining a scale type are those that keep intact the empirical information depicted by the scale.... That indeed is the principle of invariance that lies at the heart of the conception. (Stevens, 1968, pp. 103–104)

The transformations preserving the empirical structure are performed only on the numeric scale, and need have no analogue on the empirical objects.

For nominal scales, the only empirical operation is "Determination of equality." Stevens does not define equality, except by its consequence that equal objects are assigned the same numeral. He distinguishes two types of equality relations: Those that seek to identify individuals and those that partition into classes. An equality relation that distinguishes individuals can perhaps better be termed an *identity* relation, while one that denotes membership in a shared class might be termed *equivalence*. For nominal scales, any one-to-one substitution preserves the empirical structure, and thus is an allowable transformation.

The second scale type, ordinal scales, includes determination of equality, and adds a required empirical operation which Stevens refers to as "Determination of greater or less." As with equality, Stevens does not define the order relations or give any of their properties. He also does not explicitly state that the object with the "greater" of a given aspect should always be assigned a greater number, but it can be inferred from his comments. Group structure is preserved by any monotonic increasing function.

Stevens then introduces two scale types that he considers "quantitative." The first type, interval scales, again involves determinations of equality and order, and adds the operation of "Determination of equality of intervals or differences." This implies that between any two objects or attributes there exists some kind of difference. I will refer to a number of kinds of differences between attributes as *empirical differentials*, of which Stevens' "intervals" will be one type. Stevens does not require an operation to determine the size of the interval differential, or any other properties thereof, but rather requires only that there be a way to determine *equality of the intervals* themselves. This equality is intended to be isomorphic to equality of subtraction in the corresponding numeric values. This means that the empirical differential between A and B is equal to the empirical differential between C and D if and only if $f(A) - f(B) = f(C) - f(D)$.

This isomorphism is preserved by any linear transformation, but Stevens further requires that each scale type preserves the isomorphisms of the previous types, which means that only linear transformations with positive slopes are allowable, in order to preserve consistent "determination of greater or less" from ordinal scales. Figure 1.1 illustrates a representation of points on a line as an interval scale with an equal interval determination.

For the final scale type, ratio scales, Stevens identifies an additional empirical operation, namely, "Determination of equality of ratios." For this operation, equal ratios of attributes correspond to equal quotients within the numeral series. Since "ratios" are also a way to describe differences between attributes, I will also consider this in the category of "empirical differential." Note that these ratios should be taken between the attributes themselves. In Figure 1.1, the *distances* between locations can be represented as a ratio scale attribute ("length"; see Figure 1.2), but the points themselves do not lend themselves to such compar-

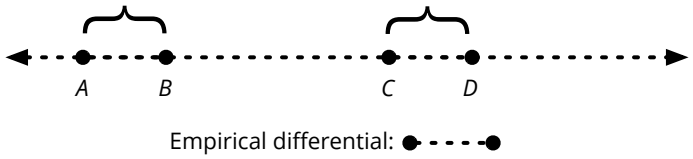


Figure 1.1: "Equality of intervals," using locations on a line as the attribute. The distance between A and B is equal to the distance between C and D.

isons. By contrast, for a property such as "area," the attributes can be directly compared using ratios (Figure 1.3).

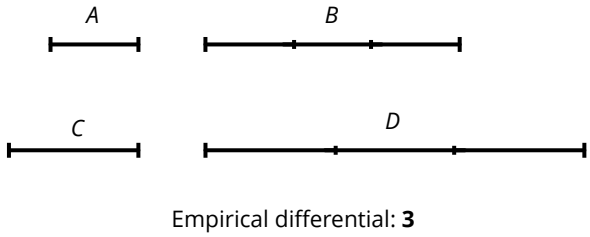


Figure 1.2: "Equality of ratios" using length as the attribute. The ratio of length between B and A is the same as between D and C (a ratio of 3 to 1).

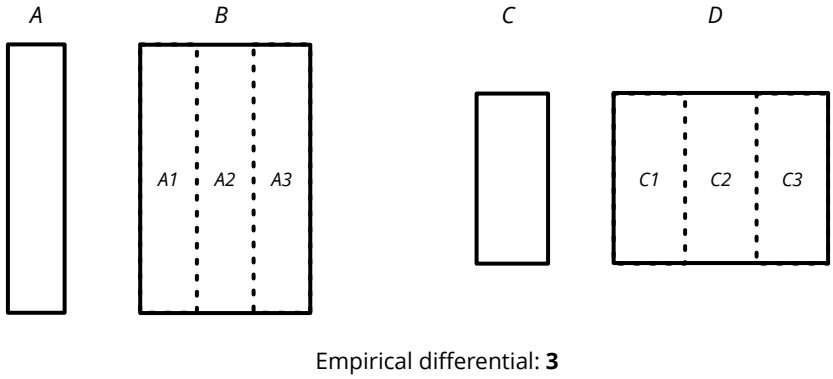


Figure 1.3: "Equality of ratios" using area as the attribute. The ratio of area between B and A is the same as between D and C (a ratio of 3 to 1).

Stevens' scale typology requires that allowable transformations on ratio scales preserve order and equal intervals as well as equal ratios, which restricts the possible functions that can be applied. Otherwise, raising to any odd integer power (or taking any odd root) could be allowed, as $\frac{a}{b} = \frac{c}{d}$ if and only if $\frac{a^{2n+1}}{b^{2n+1}} = \frac{c^{2n+1}}{d^{2n+1}}$ (for non-zero b, d and integer n). However, the

requirement that intervals be preserved as well as well forces a restriction to simple positive multiplication, which preserves not only equality of ratios, but the ratios themselves.

Also preserved by multiplication is the assignment of a zero value, leading to the designation of the quantity assigned this value to be an *absolute zero*. Additionally, if there is a meaningful empirical additive relation, it is preserved through this multiplication by the distributive law ($a + b = c$ if and only if $ka + kb = kc$). Stevens, however, does not require any kind of empirical additivity for applying a ratio scale.

Figure 1.4 depicts the four Stevens scale types and the determinations that apply to each scale type. As shown, the nominal scale has only the "determination of equality" relation, while the ratio scale has all four determinations described by Stevens.

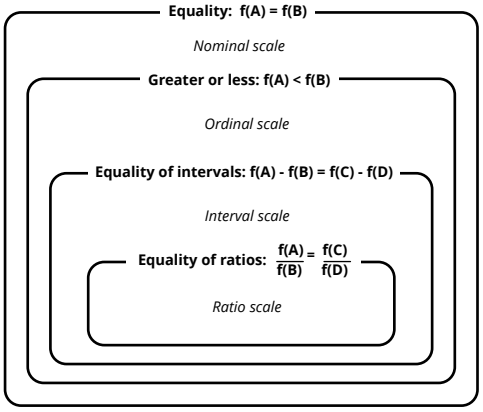


Figure 1.4: The four Stevens scales and the determinations applicable to each. Only determination of equality applies to the nominal scale, while all four determinations apply to the ratio scale.

1.3 Difference relations

1.3.1 Terminology

In order to apply the Stevens scales to psychometric models, it is necessary to formalize (and potentially extend) the typology. An alternate classification scheme, based on Stevens' system of empirical and mathematical determinations is presented below. The design of this scheme has two motivations. First, as discussed above, Stevens himself considered the determinations to be fundamental to the concept of the scale, and in following this framework I am attempting to respect the emphasis he placed on it. Second, focusing on the real-world nature of the empirical operations helps ground the typology and connect the scales to the actual objects and attributes being measured.

I will begin by defining a number of related concepts. The first three are empirical and relate to the attributes directly:

- The attributes of two objects can be empirically *compared* in some way. I will refer to this as the **empirical difference relation**. The empirical difference relations defined by Stevens are *interval* and *ratio*. The difference relation can be thought of as a question, such as "How far apart are the locations of points A and B?" or "How many times bigger is the area of rectangle A than the area of rectangle B?"

- The answer to this question can take many forms. For example, answers to the above questions may be "three inches" or "three times." This answer is what I have been calling the **empirical differential**. The empirical differential between two attributes can be a number (in the "three times" case) or a magnitude (in the "three inches" case), depending on the empirical difference relation used to compare them. While these two types of differences may seem quite dissimilar in physical examples, in psychometrics it may not always be clear which type of empirical differential is present. Additionally, as will be discussed below, the distinction Stevens makes between "ratios" and "intervals" is not necessarily a hard line. For these reasons, I believe the umbrella term of "empirical differential" can be useful to be able to talk generally about ways to compare attributes.

- Per Stevens' scale definitions, there must be some empirical operation which is held to be isomorphic to a mathematical operation. I will continue to refer to these as the **empirical determinations**. For Stevens' interval and ratio scales, the empirical determinations consist of determinations of *equality of the empirical differentials*.

In Figure 1.3, the empirical difference relation is how many more times bigger one area is than the other. The empirical differential illustrated is 3. The empirical determination is that A is as many times bigger than B as C is bigger than D.

Corresponding concepts exist on the mathematical side, and are applied to the numbers representing the objects or attributes:

- First, as an analogue to the empirical difference relations, **mathematical difference operations** may include subtraction or division.

- The **mathematical difference value** is the result of the difference operation, corresponding to the empirical differential.

- Finally, the **mathematical determination** is the operation that is isomorphic to the empirical determination.

These concepts are summarized in Table 1.1.

1.3.2 Distinguishing relations

Using this terminology, Stevens' scales involve two different empirical difference relations, called *interval* and *ratio*. In order to determine whether these two relations are evident

Concept	Empirical side	Mathematical side	Mathematical examples
Comparison between A and B	Difference relation	Difference operation	Subtraction, division
Difference between A and B	Empirical differential	Difference value	Subtractive difference, quotient
Determination of equality of differences	Empirical determination	Mathematical determination	Equality of subtraction, equality of division

Table 1.1: Related empirical and mathematical concepts

in the Rasch and 2PL models, it is necessary to understand the difference between the two relations. Both involve comparisons between two objects or attributes. In the Stevens scales, isomorphisms are established between determination of equality of intervals and equality of subtraction on the one hand, and determination of equality of ratios and equality of division on the other. This suggests one possible strategy for identifying the relations: If an isomorphism exists between the determination of equality and equality of subtraction, the determination must have been of equality of intervals; if instead there is an isomorphism to equality of division, it must be an empirical determination of equality of ratios.

Unfortunately for this strategy, it is not possible for an empirical relation to be isomorphic to only one of these two mathematical determinations, due to the relationship between subtraction and division under logarithmic functions. See Theorem 1 for a proof that if an assignment of numbers to attributes exists for which the determination is isomorphic to equality of division, then an assignment consisting of the log of these numbers will establish an isomorphism between the determination and equality of subtraction (all proofs in Appendix). The converse is also true; if an assignment induces an isomorphism to equality of subtraction, then its exponentiation will create an assignment with an isomorphism to equality of division. In other words, if a difference relation exists which is isomorphic to equality of division under one measurement map, then there exists another measurement map under which the relation is isomorphic to equality of subtraction, and vice versa.

This means that "being isomorphic to equality of division under some map" and "being isomorphic to equality of subtraction under some map" are equivalent properties, and thus cannot be used to identify the empirical relations. This equivalence is why I use the umbrella term "empirical differential" to refer to both what Stevens calls "intervals" and what he calls "ratios": The separation between the two is not as clear as his terms imply. The question then is whether there are two distinct types of empirical difference relations and empirical differentials, and if so, what could be used to distinguish the two.

Some possible approaches include:

1. **Distributivity.** If two different difference relations are present, then if the usual distributive property applies, one should map naturally to the mathematical difference operation of subtraction, while the other maps to division.

2. **Concatenation.** Some difference operations are accompanied by an empirical *concatenation operation*, which is isomorphic to addition under the measurement map and implies the presence of a true zero, while others are not.

3. **Nature of empirical differential.** Distinguish the two difference relations by the nature of the empirical differential. As discussed above, the empirical differential can be a scalar number, lacking units, as for example when the difference relation is "How many times bigger is rectangle A than rectangle B?" (e.g., "three times"). If, however, the difference relation is "How much longer is line A than line B, then the empirical differential can be expressed as a *magnitude* (e.g., "three inches"). By "magnitude," I mean a quantity such as length or mass whose value is only expressible using units. [1]

These three criteria provide different starting points for thinking about difference relations. The first criterion is closest to the presentation in Stevens (1946), in which a ratio scale is defined by the presence of both the equality of ratios empirical determination and the equality of intervals empirical determination. This can be illustrated using area, as different areas can be empirically compared either as three additional square inches, or as three times as large. These two area difference relations are illustrated in Figure 1.5.

Both panels of Figure 1.5 show valid representations of a relation for which "B is as much bigger than A as D is bigger than C." In the upper panel, this is true when comparing the ratios of the areas of the respective shapes. In the lower panel, the equality holds when considering the amount of extra area the larger shapes have relative to the smaller shapes.

I have named this criterion "Distributivity" to indicate its reliance on the distributive property of arithmetic. When only one difference relation is present, isomorphic to either subtraction or division, it cannot be classified using this criterion. Stevens suggests that it be treated as subtractive, but in fact it is perfectly possible to define a scale with a single difference relation which is isomorphic to division under the map to the reals. This scale is not included in the Stevens typology, but will play a part in the discussion of the scale types of the IRT models.

Using the second criterion, *concatenation*, the "ratio" difference relation is defined when objects or attributes can be physically added together to yield another element. This is true of areas (Figure 1.6). By contrast, locations cannot be concatenated in a meaningful way. This criterion is useful for attributes such as length and mass in which concatenation comes naturally.

There is a long history in measurement of considering an empirical concatenation operation to be a fundamental property of a measurable quantity. Campbell and Jeffreys (1938),

[1] The third chapter of this dissertation discusses the concept of magnitude in more detail. For now, it is enough to distinguish magnitudes, which have units, from raw, scalar numbers.

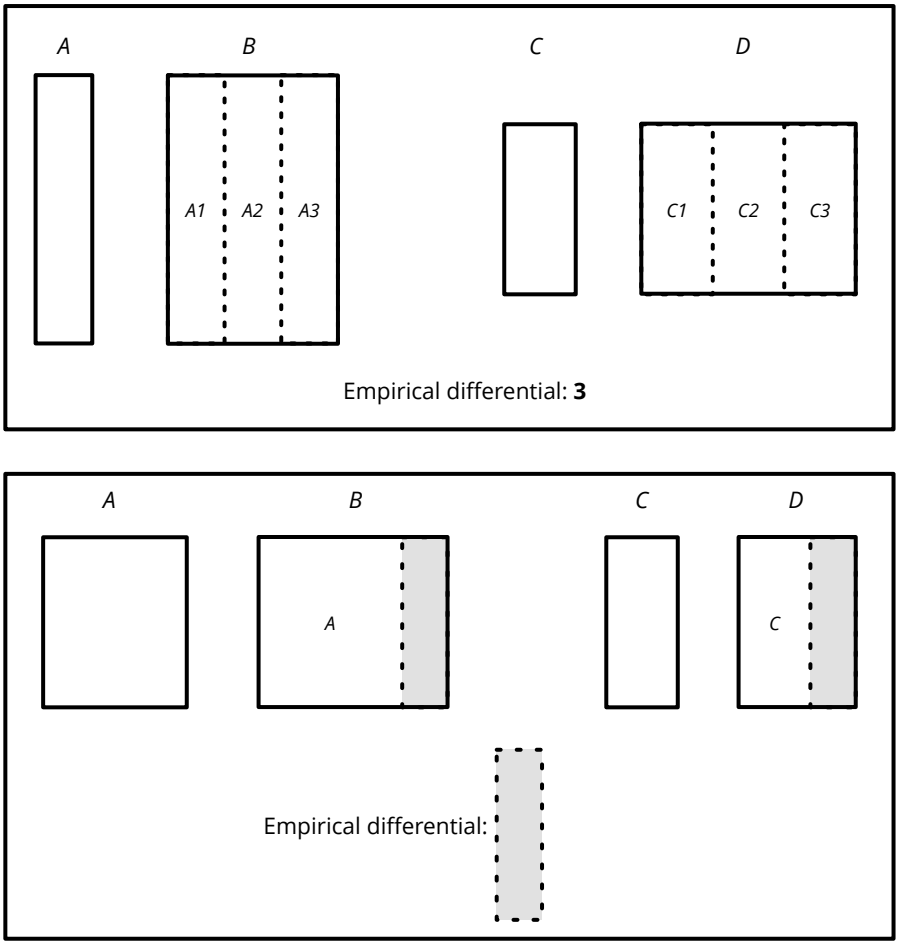


Figure 1.5: Two empirical difference relations for area, with corresponding empirical differentials.

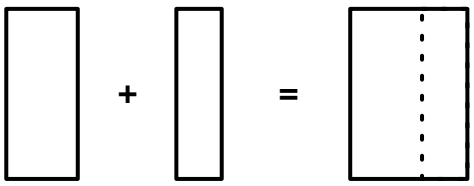


Figure 1.6: Concatenation of two areas.

for example, described additivity of an attribute, defined as the presence of "a general operation by means of which two systems can be combined (in some very general sense) so as to produce a third greater than either of them," as a necessary condition of fundamental measurement. Hölder (1901) included concatenation in his first set of axioms of quantity defining a magnitude (Michell & Ernst, 1996).

I have labeled the last criterion *Nature of the empirical differential*. Again, the empirical differential is the result of the empirical difference relation, and is what is compared in the empirical determination. Figures 1.7 and 1.8 present two different types of empirical differentials.

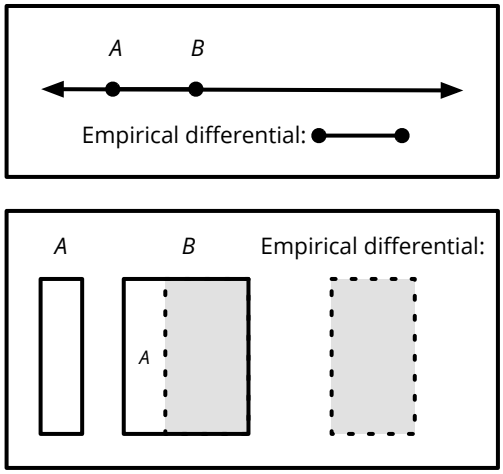


Figure 1.7: Difference relations for which the empirical differentials are magnitudes.

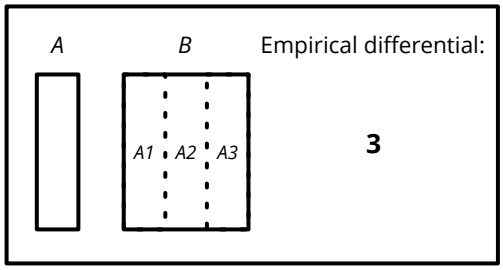


Figure 1.8: A difference relation for which the empirical differential is a scalar value.

The first type of empirical differential, shown in Figure 1.7, is what I am calling *magnitudes*. This type includes the empirical differential for location, which consists of a distance, with a certain length. For area, the empirical differential for the interval difference relation is itself a region, with a certain area. For both of these interval relations, the size of the empirical differential can only be expressed using units: three inches farther, or three more acres. In contrast, Figure 1.8 shows a difference relation with a scalar empirical differential.

The area of B is three times the area of A, for a empirical differential of 3 (expressible without units).

Using the "Nature of the empirical differential" criterion permits a focus on the relationship between the empirical differentials and the mathematical difference values. In Stevens' scale definitions, the mathematical determination of equality is isomorphic to the empirical determination of equality. This means that there must be a bijection between the empirical differentials and the mathematical difference values (which I will call the *difference bijection*).[2] If difference relation has a scalar empirical differential, meaning that the empirical differential is a real number, it is sometimes possible for the difference bijection to be the identity function. If this occurs, or in some cases if the bijection is another continuous, selected function (such as, for example, *log*), then I will refer to the scale as an *absolute* scale, with the alternative being a *relative scale*.

"Absolute" here indicates that the mathematical difference values are fixed and can be determined using only the empirical differentials. For a ratio scale, the ratios are predetermined, so the transformations that preserve the group structure are scalar multiplications, marking this as having the same structure as Stevens' ratio scale definition. The difference bijection in most ratio scales is typically the identity function. This means the empirical differential between two objects is a real number in these cases, and that number is equal to the mathematical ratio between the numeric values assigned to the objects. When this situation occurs, the measurement process is often referred to as involving "finding the ratio" between two objects. For example, Michell (2005) writes:

> If Smith's weight is 90 kg, then this is equivalent to asserting that the real number, 90, is a kind of relation, viz., the kind of relation holding between Smith's weight and the weight of the standard kilogram.... This position entails that measurement is the attempt to estimate the ratio between two instances of a quantitative attribute, the first being the magnitude measured, and the second being a known unit.

In contrast, it is not true for interval scales that the mathematical difference value is uniquely determined by properties of the empirical differential alone. In Figure 1.7, if an interval scale is applied to these locations with numerical values $f(A), f(B) \in \mathbb{R}$ assigned to A and B respectively, the difference value $f(B) - f(A)$ may be any real number, marking this as a relative scale. Compare this result to Figure 1.8, where any normal ratio scale assignment will ensure that $\frac{B}{A} = 3$, making this a difference value which is not only determined by the empirical differential, but is in fact equal to it. Common relative scales include location markers, temperature scales without absolute 0, and many attitudinal scales.

[2] A *bijection* is a mathematical function f from a set X to Y such that for every $y \in Y$ there is exactly one $x \in X$ such that $f(x) = y$. If there is an isomorphism between the determinations of equality, then two mathematical difference values will be equal only if their corresponding empirical difference values are equal, establishing the bijection.

In this paper I will use the "Nature of the empirical differential" criterion to distinguish difference relations. The interval scale is then defined as a relative scale with a subtraction difference operation, while a ratio scale is an absolute scale with a ratio difference operation. I will refer to the scale typology derived from this distinction as the "empirical differential isomorphism" (EDI) typology.

1.3.3 Non-Stevens scale types

Complementing the interval and ratio scales are two more, not discussed by Stevens but relevant to the purposes of this paper. Within the framework of the EDI typology, the first is a scale in which a scalar difference is represented by subtraction, and the second is a scale in which a magnitude difference is represented by division. The first scale type was referred to as simply a "difference scale" by Suppes and Zinnes (1963), a term they credit to Donald Davidson, but I will refer to is as an *absolute difference scale* for clarity. The absolute difference scale, like the interval scale, has the property that equality of empirical differentials is isomorphic to equality of subtraction, but it also has the further property that the subtractive difference itself is fixed and determinable from the empirical differential (either equal to it or a simple, pre-selected transformation).

Suitable examples of absolute difference scales are not readily available. There exist two common types of empirical operations producing scalar values. One type, *ratios*, has already been discussed, and maps more naturally onto division. The other is counting. Figure 1.9 illustrates an absolute difference relation built upon count differences. The limitation in this example is that the counting operation is typically limited to positive integer results, while the absolute difference scale allows for any real difference value.

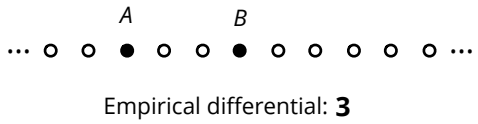


Figure 1.9: An absolute difference scale, in which the empirical differential between two elements can be determined by counting the elements in between.

Another example of absolute difference scales comes from games where the object is to have the most points at the end. A game of Hearts in which the scores of the four players are 0, 4, 15, and 33 points is essentially in the same state as one in which the scores are all 13 more at 13, 17, 28, and 46; what is meaningful is the difference between the scores.

Finally, absolute difference scales can be constructed from ratio scales by taking the empirical difference operation to be the log of the empirical ratio operation. This is perhaps a more suitable operation in that it readily produces any real number difference value, but its required mathematical log operation prevents it from being fully satisfactory.

The second new scale has the property that equality of empirical differentials is isomorphic to equality of ratios (like the ratio scale). Unlike the ratio scale, it does not involve a set

value for the ratio between two objects. This has been called a "logarithmic interval" scale (Stevens, 1957) or "log-interval" scale (Narens & Luce, 1986). I will refer to it as a "relative ratio" scale. One possible way to visualize this scale type is to imagine labeling an infinite slide rule. A slide rule has the property that any pairs of points that are the same distance apart should be labeled with numbers that are the same ratio apart. In Figure 1.10, the top slide rule shows unlabeled points A, B, C, D. Due to their relative placements, the ratio between the values assigned to A and B should be the same as the ratio between the values assigned to C and D. The two slide rules below it show possible valid labelings that preserve this property, as well as some other labeled intermediate points. Within each slide rule, any points whose labels are in a 1:2 ratio are the same distance apart from each other, although this distance is smaller on the bottom slide rule than the middle rule.

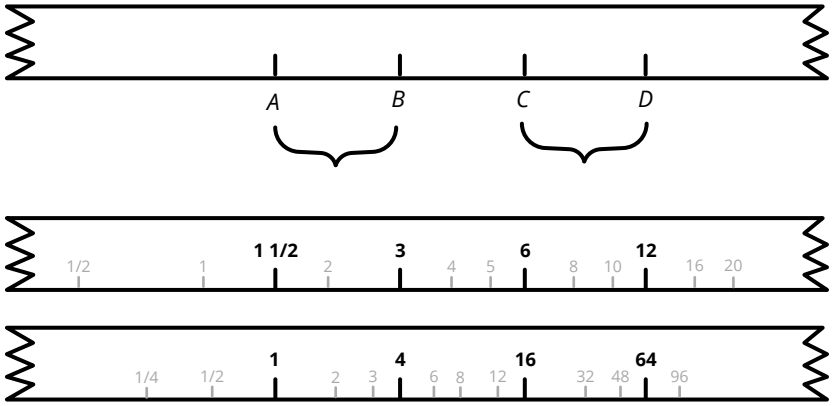


Figure 1.10: A relative ratio relation. On this infinite slide rule, the ratio between the values assigned to A and B should be the same as the ratio between the values assigned to C and D.

For consistency, we might refer to the original ratio and interval scales as "absolute ratio" and "relative difference" scales, respectively. Table 1.2 describes the four scale types in terms of their empirical differentials and corresponding mathematical difference relations.

Empirical differential is...	Isomorphic to quotient	Isomorphic to subtractive difference
A real number	Ratio	Absolute difference
A magnitude	Relative ratio	Interval

Table 1.2: Scale types by empirical differential type and mathematical difference operation. Again, I am using the term "magnitude" to refer to quantities whose values are expressed with units.

These two novel scales can be constructed from the two original scales. Any property which can be expressed using a ratio scale can be represented on an absolute difference scale

in which the assigned numeric values are logs of the ratio scale values (Theorem 2). Similarly, any property represented on an interval scale can be described in terms of a relative ratio scale by exponentiating the numeric values assigned under the interval scale.

1.4 The IRT models

1.4.1 Rasch model paradigms

In order to apply the EDI typology to IRT models, it is necessary to define what constitutes a "difference relation" between two respondents. In the context of achievement tests, the difference between two respondents' aptitudes can be thought of as how much better Respondent A is than Respondent B at the skill, how much more knowledge Respondent A has than Respondent B in the area, etc. In terms of the previously defined terminology, this empirical differential seems to lend itself to relative scales rather than absolute scales, since "How much better is Alice than Bob at math" does not seem like the type of question that can be answered by a scalar number.

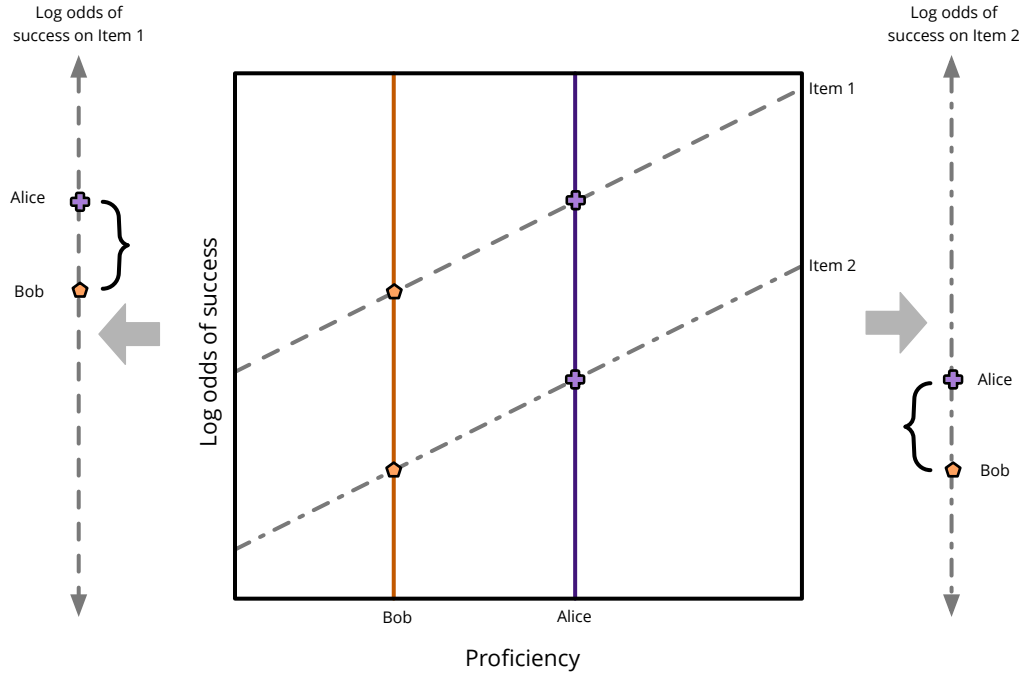
However, under a Rasch model or Generalized Rasch Model, there is a specific sense in which statements such as "Alice is twice as proficient as Bob" can be understood meaningfully. This sense relates to Alice's *odds of success* on an item. When the item discriminations are equal, the ratio between Alice's odds of success on an item and Bob's odds of success on an item will be a constant across all items, independent of the item's difficulty (Theorem 4). Equivalently, the difference in log odds of success for Alice and Bob is constant across items (Figure 1.11a).

This relationship meets the requirement specified by Rasch for a meaningful comparison:

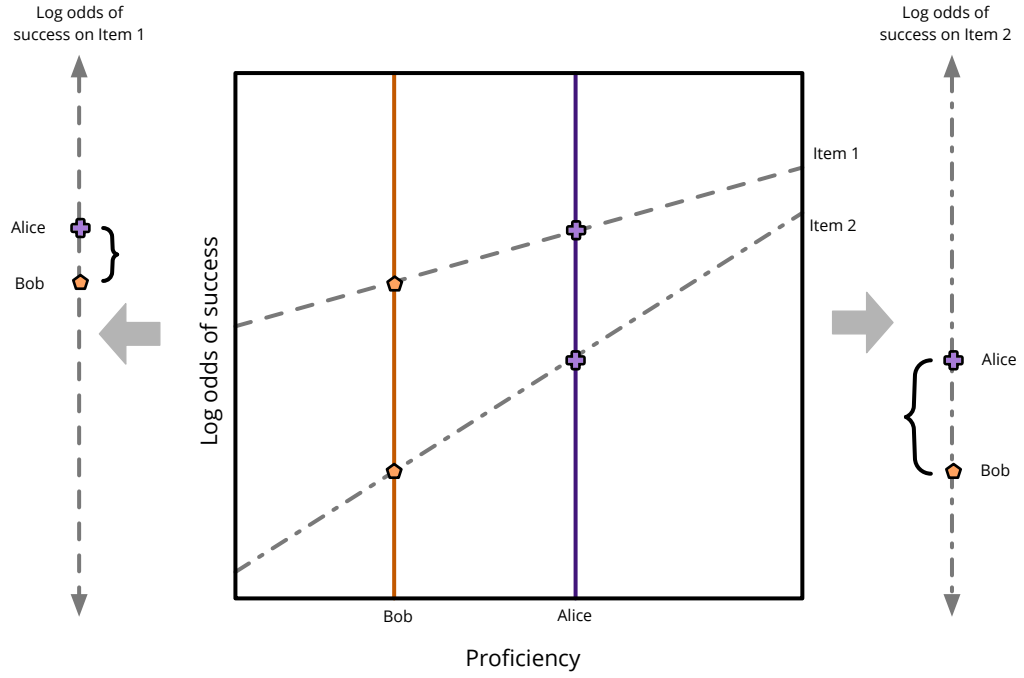
> If the statement that the ability of one person is twice the ability of another person... shall be of any use, it must be valid in connection with more than one problem. It must remain in force *when we present the persons with several problems of the same kind.* (Rasch, 1960/1980, p.72, emphasis original.)

The value of this odds ratio can then be interpreted as an empirical, real-valued difference between Alice and Bob. However, this interpretation comes with a key caveat. Namely, it does not extend to an item set with a different discrimination. On those items, the ratio between Alice's predicted odds of success and Bob's predicted odds of success will be different, as will be, equivalently, the difference in their log odds of success (Figure 1.11b). Of course, for a set of items that conform to the Rasch model, there are no items with different discriminations. The meaning and implications of this proscription are crucial to determining the model's scale type.

During the assessment design process, test developers using the Rasch model to guide test development routinely discover, and subsequently refine or eliminate, items with variant discriminations. Under one point of view, constructing an item set to fit the Rasch model can be seen as analogous to trying to construct a ruler with evenly spaced tick marks (Andrich,



(a) In the Rasch model, the difference in log odds of success for two respondents is constant across items.



(b) The difference in log odds of success for two respondents is different on items with different discriminations.

Figure 1.11: Chances of success for two respondents on different items in a Rasch and 2PL model.

2004). There is no inherent benefit to using inches over centimeters or vice versa, or analogously to using one or another set of items; what is important is the consistency of the unit within the set. To further stretch this metaphor, under this understanding the statement "Alice is twice as proficient as Bob" is no more meaningful than "Alice is two (units) taller than Bob." The statement cannot be fully generalized beyond the measurement instrument (other than to equally discriminating items outside of the instrument), and thus does not provide sufficient meaningful information about the trait.

An alternate understanding, however, holds that an alternate discrimination is an indication of a serious issue with the item. Such an issue not only warrants its removal from the particular assessment for which its presence induces a misfitting model, but excludes it from the universe of acceptable items for measuring the given construct. The problem is not that it is an "inch" item on a "centimeter" ruler; the problem is that it is measuring something other than height. This "something else" that variantly discriminating items are capturing has been suggested to be placement in an assessment (Yen, 1980), test-wiseness or fatigue (Masters, 1988), noise (Ferrando, 2009), a slightly different skill (Samejima, 1969), or simply another, related construct (Reckase & McKinley, 1991). Under these points of view, the fact that Alice's odds are not twice Bob's on items with different discriminations is understood as consistent with the fact that Alice is not twice as proficient as Bob on *a different construct*. For the construct being measured, for which all items discriminate equally, the statement is meaningful.

I will refer to these as the "selected slope" and "ideal slope" paradigms for the Rasch model. Under the selected slope paradigm, a group of items with the same discrimination is selected or constructed in order to fit a Rasch model, but other valid sets of items measuring the construct, with alternate slopes, are assumed to exist. Empirical relations on the attribute should therefore be true not just within the constructed set, but for the variantly discriminating items as well. Alternately, within the ideal slope paradigm, it is assumed that there is one "ideal" discrimination for a given construct, and thus empirical relations only need to hold for items with that true slope to be meaningful.

The process of Rasch instrument design does not differ under the two paradigms. In both cases, a set of items is constructed, analyzed, and adjusted, with the goal of obtaining a set of items with roughly equal discriminations. The only difference is philosophical: Is the common item discrimination assumed to be the "true" discrimination for that construct, or is it assumed that a number of other, equally valid item sets could have been constructed for this construct, each with its own discrimination?[3]

The ideal slope paradigm is similar to the idea behind general objectivity. Specific objectivity refers to the idea that comparisons between person abilities ought to be independent

[3]When I refer to "slopes" or "discriminations" in this context, I am referring to the rate of change of log odds of success as a function of person proficiency. I am *not* referring to the numeric parameter value assigned for this purpose in the model (typically, 1). The ideal slope paradigm does not imply that there is a true discrimination parameter value of 1 or 3.5 or any other number. It states only that for any construct, there is a true slope that ought to be found within any valid item measuring that construct, but which can be assigned any numeric value.

of the items, while comparisons between item difficulties ought to be independent of the respondents (Rasch, 1966). This concept can be split into the notions of *local* and *general* objectivity (Stenner, 1996). Local objectivity is simply an empirical property of data that fits the Rasch model. However, general objectivity is a theoretical conviction that measurements will be independent of any instrument used (Stenner, 1994). Local objectivity therefore is the result of an instrument constructed under either paradigm, while belief in the presence of general objectivity corresponds to belief in the ideal slope paradigm.

While compromise positions between the two main paradigms are imaginable, they can generally be identified with one or the other option. For example, one possibility is that items measuring a construct have a true *mean* discrimination, but with some amount of variance around that mean. Constructing a Rasch item set would then involve eliminating or modifying items whose discriminations fell too far from the mean discrimination, and accepting the small variance within the final set as an acceptable approximation to constant variance. Under this paradigm, "Alice is twice as proficient as Bob" would then be meaningful relative to the population mean. I classify this compromise as a subtype of the "ideal slope" paradigm, as there is a sense in which the ratio comparisons are meaningful, and the selection of common discrimination for the item sets is not treated as arbitrary.

My primary goal in introducing these contrasting paradigms is to clarify why and how the Rasch model can be classified as multiple scale types, under different analyses. These analyses necessarily depend, explicitly or implicitly, on whether the common slope present in a set of Rasch items is thought to be arbitrary or inherent to the construct.

1.4.2 Scale types for proficiencies

Under the "selected slope" paradigm, proficiencies under the 2PL and Rasch models will have the same scale type, since both involve the same extended universe of possible items. The Rasch model just requires using a more specifically curated set of such items in any given assessment. The fact that the Rasch model has desirable statistical properties not found in the 2PL model (specific objectivity, sum score as sufficient statistic, double monotonicity, etc.) does not mean that proficiencies are measured along a different type of scale.

Under the "ideal slope" paradigm, however, the Rasch model is making different assumptions from the 2PL model. These assumptions concern not only the specific items within any one instrument, but how performance on the construct behaves in general. For this reason, it can have a different scale type from the 2PL model.

Within the ideal slope paradigm, "Alice is twice as proficient as Bob at this construct" is meaningful, since all items for which she does not have twice his odds of success are not considered to validly measure the construct. This means that there is a real scalar number (2) serving as the empirical differential between Alice and Bob. For the odds form of the Rasch model, this empirical differential will be the same as the mathematical ratio between their proficiency parameters (Theorem 5), making this form of the model a ratio scale with an identity difference bijection. For a GRM with a non-unitary discrimination index, the mathematical ratio will be equal to the empirical differential, raised to the power of the

reciprocal of the consistent discrimination index (Theorem 6). This is still structurally an absolute ratio scale, but without the identity bijection between empirical differentials and mathematical difference, making it a less natural choice for representing the relationship.

In the log odds form of the Rasch model, the subtractive difference between the parameter values will be the log of the empirical differential (Theorem 11). Since this is a fixed constant that depends only on the empirical differential, this form of the Rasch model is an absolute difference scale. This structure also holds for the log odds form of the GRM, assuming the discrimination index is fixed (Theorem 10). This gives us two possible scale types for the Rasch model under the ideal slope paradigm: absolute ratio, and absolute difference.

For a 2PL model, or within the selected slope paradigm, statements such as "Alice is twice as proficient as Bob at this construct" are not meaningful, since they would not be true had a different set of items been selected, with a different common discrimination parameter. However, it is still possible to make an empirical determination of "equality of difference." That is, there is a sense is which the statement "The difference between Alice's and Bob's proficiencies is the same as the difference between Carl's and Diane's proficiencies" is meaningful, in that it is true for all the items in the extended universe. This sense arises from comparing the odds ratios of success of the two pairs. Under a 2PL model (and the related forms of the Rasch model), if the odds ratio of success between Alice and Bob is the same as the odds ratio of success between Carl and Diane on one item, then it is the same on any item, even if that item has a different discrimination (Theorem 8). Equivalently, if the difference in log odds of success between Alice and Bob for one item is the same as the difference in log odds of success between Carl and Diane for that item, then this relationship will hold for any item (Figure 1.12).

Under the odds form of the 2PL model, if two pairs of respondents have equal odds ratios of success, then the ratios of their proficiency parameters are also equal, and vice versa (Theorem 7). This "equality of ratios" puts the 2PL model, or constructed-set Rasch model and GRM, within the ratio scale family, but as a relative ratio scale rather than the more common absolute ratio scale, since there is no real-valued empirical differential relating to this ratio. In the logit form of these models, having equal odds ratios is equivalent to equality of subtractive difference (Theorem 12). The "equality of subtractive differences" determination marks these models as relative difference (i.e. interval) scales.

Table 1.3 shows the four different scale types that can be assigned to the Rasch or 2PL models: absolute ratio, absolute difference, relative ratio, or relative difference (interval). The ratio scales correspond to the odds form of the model, while the difference scales represent the logit form of the model. The absolute scale types only apply if we believe that the only possible items validly measuring the construct must have the same discrimination as each other, guaranteeing fixed odds ratios of success between respondents. Otherwise, the relative scale types are appropriate. For a 2PL model, only the relative scale types apply.

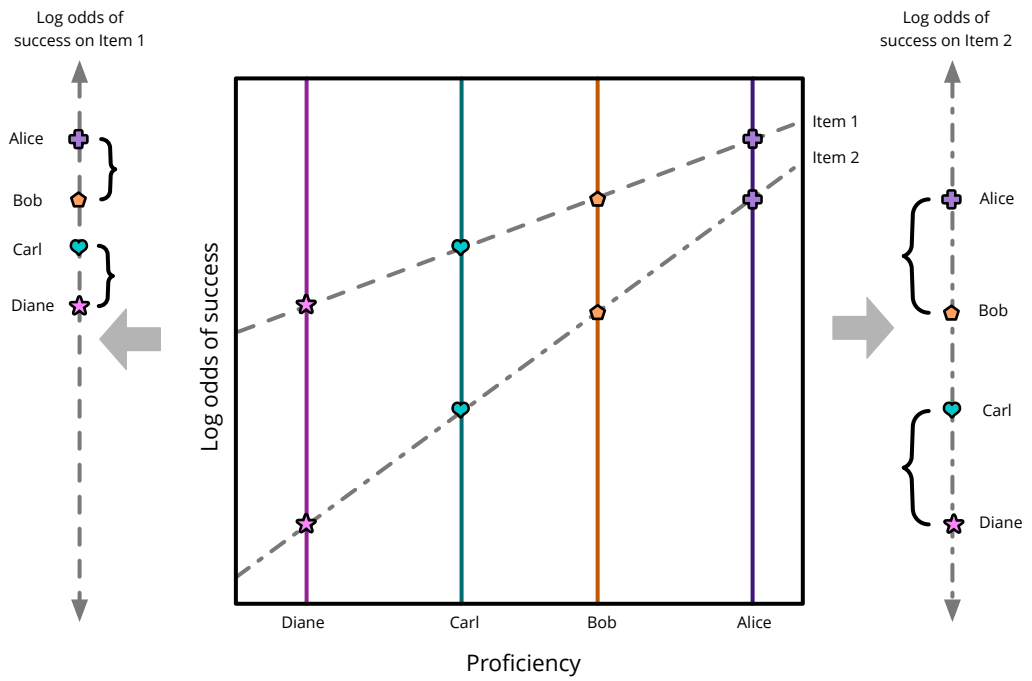


Figure 1.12: The difference in log odds of success between Alice and Bob is the same as the difference in log odds of success between Carl and Diane. This relationship holds across all items.

	Odds form	Logit form
Ideal slope Rasch	Ratio	Absolute difference
Selected slope Rasch	Relative ratio	Interval
2PL model	Relative ratio	Interval

Table 1.3: Scale types of the Rasch and 2PL models by paradigm and model formulation.

1.4.3 Scale types for items

Thus far, I have only discussed scale type properties of the person parameters. Similar analyses can be applied to the difficulty parameters of the items.

The simplest case is the ideal slope Rasch model. In this model, the item and person parameters function symmetrically. Whereas on the person side, the odds ratio of success between two respondents on a single item was constant regardless of item difficulty, on the item side the odds ratio of success of one person on two items is constant regardless of person ability (Theorem 14). Equivalently, the difference in log odds of success between two items is constant (Figure 1.13).

In the odds form of simple Rasch model with all the discrimination parameters set to 1, the ratio of the difficulty parameters of two items will be the reciprocal of the ratio

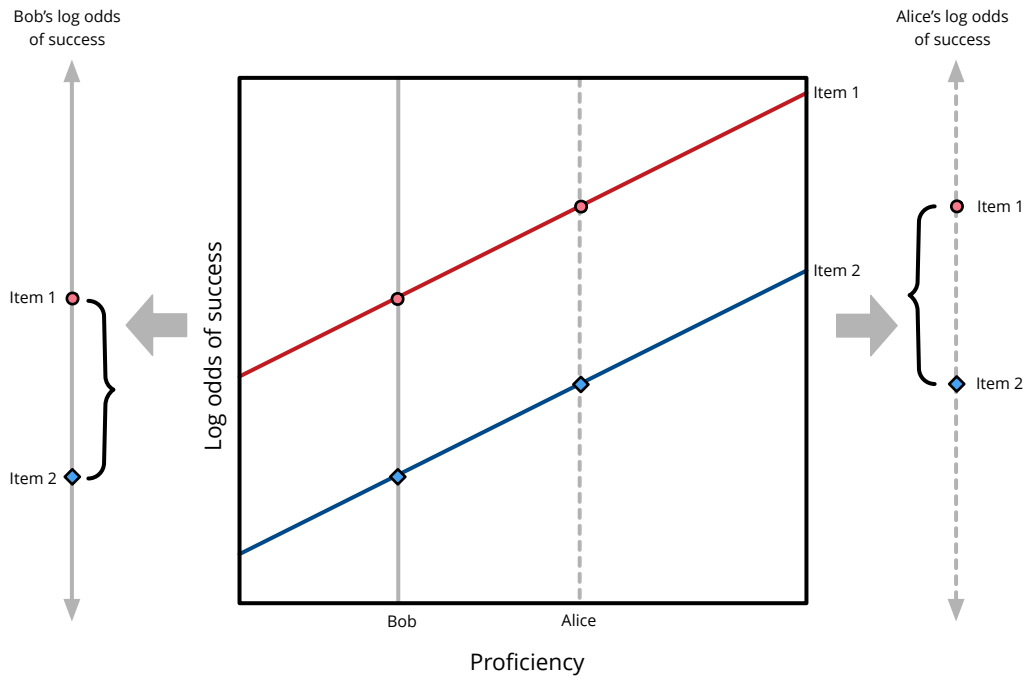


Figure 1.13: In the Rasch model, the difference in log odds of success for two items is constant across respondents.

between respondents' odds of success on the two items (Theorem 16). Since there is an empirical differential between two items that has a scalar value, and this value is equal to the ratio between the numerical values assigned to the items after a simple transformation (the reciprocal), this meets the requirements for an absolute ratio scale under the EDI typology.

A variant of the Rasch model exists that codes the item parameter as "easiness" rather than "difficulty" (e.g., Perline, Wright, & Wainer, 1979; Brogden, 1977). The odds version of this model is given by:

$$\text{Odds}(x_i = 1|t) = t \cdot b_i \tag{1.6}$$

In the odds form of this "easiness" model, the ratio between a respondent's odds of success on a pair of items will equal the ratio between item (easiness) parameters, rather than being reciprocals, creating a more direct ratio scale.

In the odds form of the GRM, if a non-unitary discrimination is used, the ratio between the item difficulty parameters is equal to the constant odds ratio, raised to the power of the negative reciprocal of the discrimination parameter (Theorem 15). Since there is still a scalar-valued empirical differential (the constant odds ratio), and the ratio between the values assigned to the items is a predetermined function of this ratio, this is still a ratio scale under the EDI typology.

In the log odds form of the Rasch model, the subtractive difference between two item parameters will be a function of this constant odds ratio of success of any respondent on the two items. Specifically, it will be the log of the reciprocal of this odds ratio (Theorem 19). Since the subtractive difference depends on this (scalar) constant empirical differential, this is an absolute difference scale under the EDI typology.

The 2PL case is more complicated. If the odds ratio of success between two items is again used as the empirical differential, then its value is not constant across different persons. Figure 1.14 depicts this in terms of log odds: Items 1 and 2, for example, are closer together in terms of difference in predicted log odds of success for Alice than for Bob, whereas Items 3 and 4 are much closer for Bob than for Alice.

In the proficiency case discussed in Section 1.4.2, it was also true that the value of the odds ratio (or difference in log odds) was not constant within a 2PL model (as shown in Figure 1.11b). However, in that case, while the *value* of the odds ratio/difference in log odds changed, *equality of differences* was constant. Two pairs of respondents who were the same "distance apart" on one item, in terms of odds ratio or difference in log odds, were the same difference apart on any item (illustrated in Figure 1.12). On the item side, by contrast, equality of differences does not hold. It is perfectly possible for two pairs of items to have equal odds ratios for one person, but not another (or, equivalently, equal differences in log odds, as shown in Figure 1.14).

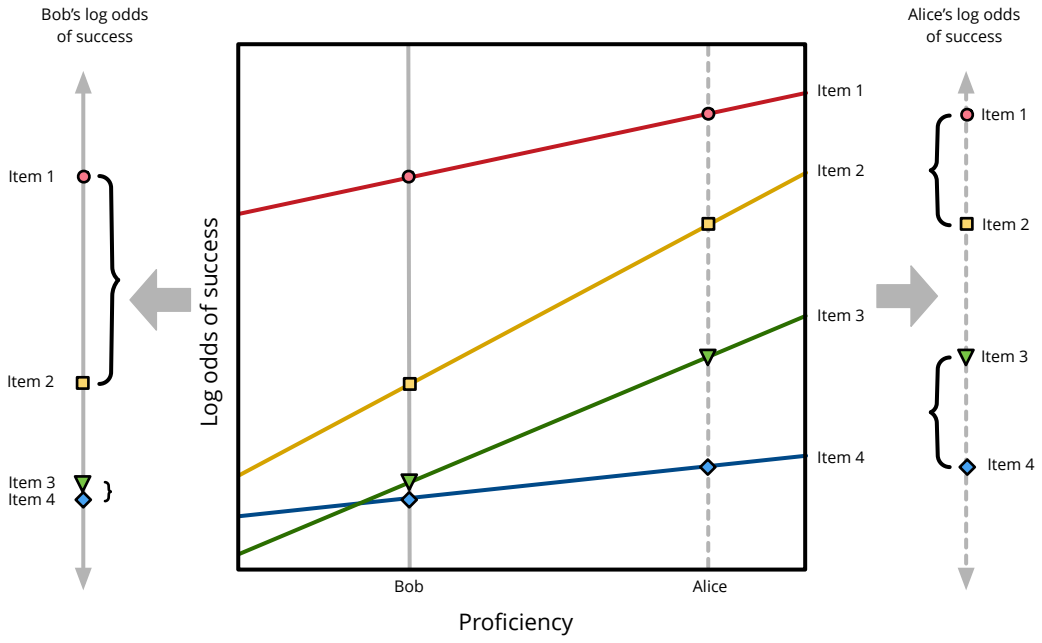


Figure 1.14: Items 1 and 2, for example, are closer together in terms of difference in predicted log odds of success for Alice than for Bob, whereas Items 3 and 4 are much closer for Bob than for Alice. Furthermore, for Alice, Items 3 and 4 are as far apart in terms of log odds as Items 1 and 2, but for Bob they are closer.

Even the order of the items can shift. Alice may have a greater predicted odds of success on Item 1, while Bob has a greater predicted odds of success on Item 2. Figure 1.15 illustrates the difference between how item ordering works in the Rasch model vs. the 2PL model. Items can be conceptualized as functions relating chance of success on an item to person proficiency. In a Rasch model, these functions do not intersect, so the items can be ordered identically by chance of success for any respondent (Figure 1.15a). In a 2PL model, the functions cross, so the item order for one respondent is different from the order for another respondent (Figure 1.15b). The item order then depends on the choice of respondent. For respondents with extreme high and low proficiencies, the item order will be exactly reversed (provided no two items have exactly the same discriminations). An extreme low respondent will order items by their discriminations such that higher discriminating items will have lower chances of success; an extreme high respondent will have the opposite order (Figure 1.16).

For these reasons, items in the 2PL model are often said to be un-orderable (Cliff, 1992), as two respondents will disagree on which one is harder. This would make 2PL items incompatible with any of Stevens' scales, even nominal scales. A nominal scale requires that two objects be at least categorizable as equal or different, but in a 2PL model, two items which have equal predicted odds of success for one respondent may not be equal for another.

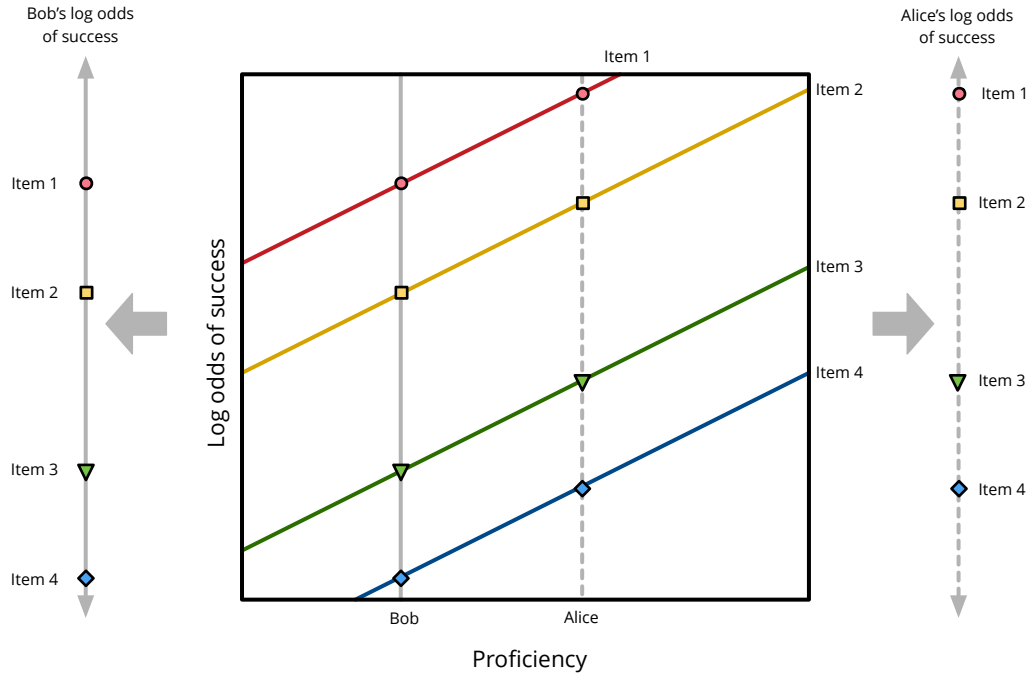
In practice, the form of the 2PL model does induce a specific ordering on the numbers assigned as its items' difficulty parameters. Consider the relationship between a difficulty parameter (b_i) for Item i and the proficiency parameter (t) of a respondent who has a 50% chance of success on an item (i.e. even odds):

$$\text{Odds}(x_i = 1 \mid t) = 1 \tag{1.7a}$$
$$\left(\frac{t}{b_i}\right)^{\alpha_i} = 1 \tag{1.7b}$$
$$\frac{t}{b_i} = 1 \tag{1.7c}$$
$$t = b_i \tag{1.7d}$$

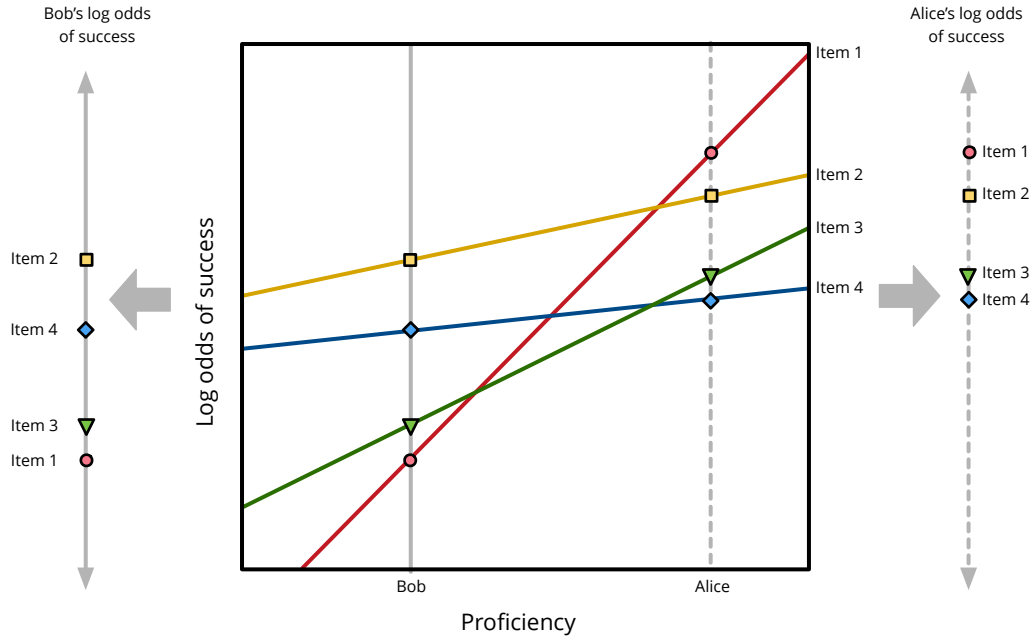
This relationship also holds true for the log odds form of the model. A respondent with a 50% chance of success on an item has a log odds of success of 0:

$$\text{logit}(x_i = 1 \mid \theta) = 0 \tag{1.8a}$$
$$\alpha_i(\theta - \beta_i) = 0 \tag{1.8b}$$
$$\theta - \beta_i = 0 \tag{1.8c}$$
$$\theta = \beta_i \tag{1.8d}$$

This means that regardless of the form of the model, the discrimination parameters, or any other constraint choices, item difficulty parameters are equal to the proficiency parameter of a respondent with a 50% predicted probability of success on that item. The use of this threshold is commonly referred to as "RP50" for "response probability of 50%" (Kolstad, 1996). Call a respondent with a 50% predicted probability of success on a specific item a *reference respondent* for that item. One item is then given a higher difficulty parameter in



(a) Items in a Rasch model have the same order for all respondents.



(b) Items in a 2PL model have different orders for different respondents.

Figure 1.15: Chances of success for two respondents on different items in a Rasch and 2PL model.

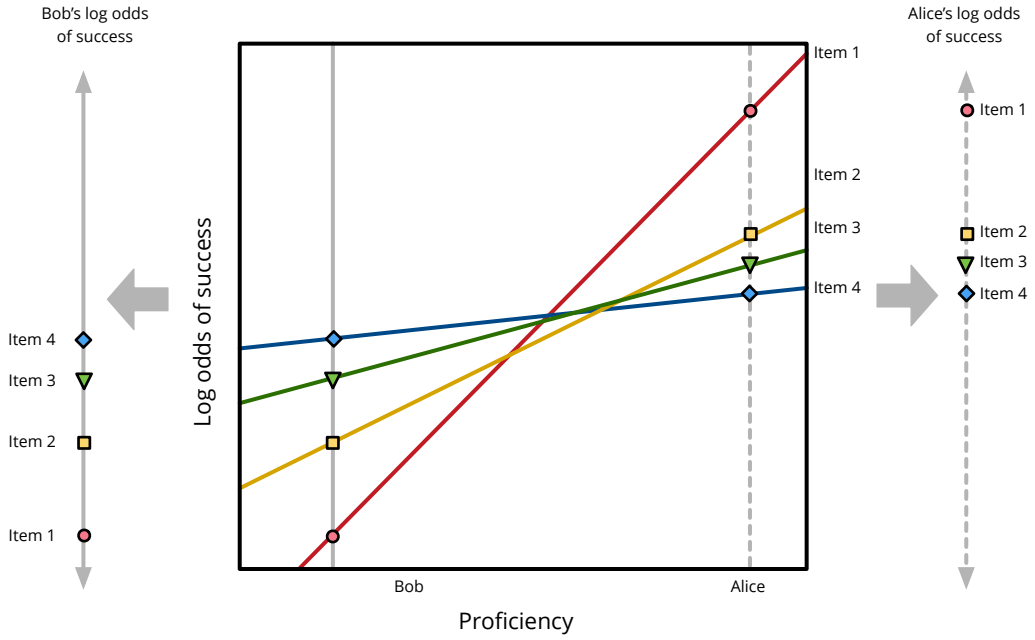


Figure 1.16: For respondents with very low and high proficiencies, item order on non-parallel items will be reversed.

the 2PL model if the proficiency of its reference respondent is higher. Note that while the order of items is not consistent across respondents in a 2PL model, the order of respondents *is* a consistent property, regardless of item discrimination (Figure 1.17). This means that there is a method to ordering items within a 2PL model: by ordering them by their reference respondents. Figure 1.18a illustrates this ordering.

This also suggests a possible empirical differential for 2PL items: The difference between the difficulties of the two items is defined as the distance between two respondents with 50% chance of success on the two items. In Figure 1.18a, the reference respondents for Items 1 and 3 are as far apart as the reference respondents for Items 3 and 4, so we could claim that their respective referent items are equal distances apart in terms of difficulty. While this definition is consistent, and suggests that the scale types for 2PL items will be the same as for respondents under the model, it is misleading. The aforementioned relationship between the reference respondents for Items 1, 3, and 4, for example, does not imply that a given respondent will find the increase in difficulty from Item 1 to Item 3 to be the same as the increase in difficulty from Item 3 to Item 4. In fact, there may be no proficiency at which the respondent's order matches the RP50 order (Figure 1.18b).

Variants of the 2PL model exist which shift the connection between items and their reference respondents. For example, if the item parameter set is scaled by a constant k relative to the person parameter set, the following model results:

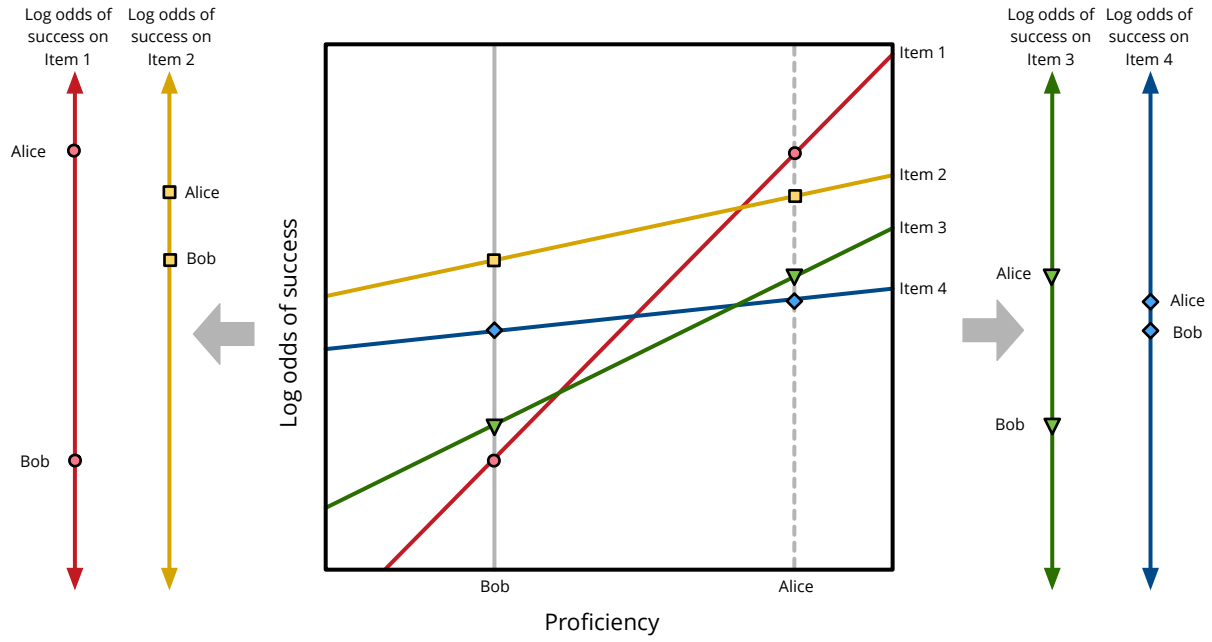


Figure 1.17: On each item, Alice's chances of success are higher than Bob's, so person order is consistent across items.

$$\text{Odds}(x_i = 1|t) = \left(\frac{k \cdot t}{b_i}\right)^{\alpha_i} \tag{1.9}$$

Or in log odds form (with $\kappa = \log k$):

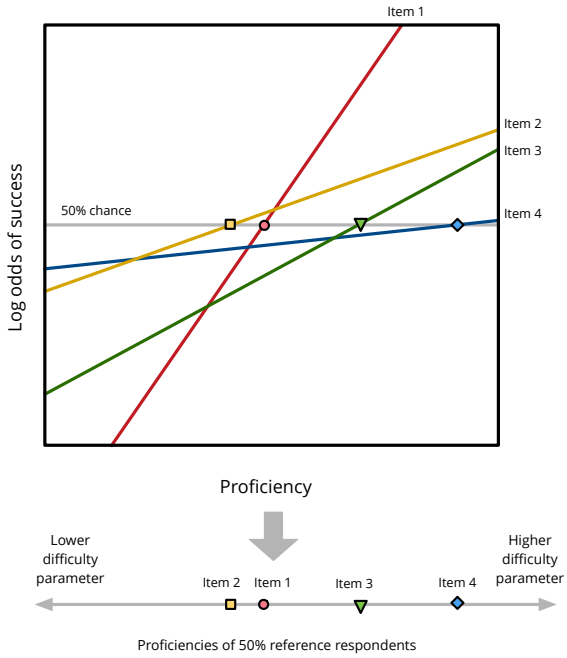
$$\text{logit}(x_i = 1|\theta) = \alpha_i(\theta + \kappa - \beta_i) \tag{1.10}$$

These variant models are equivalent to a standard 2PL model, with the same scale types. They simply adjust the relative positions of the two item sets. In these models, items are assigned difficulty parameters equal to the proficiency parameters of respondents with odds k of success on an item. This can result in different item ordering, despite no fundamental change in the model (Figure 1.19).

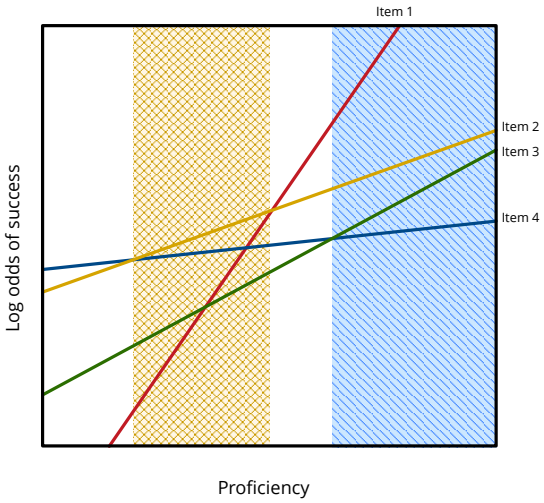
While the difficulty parameters cannot be properly ordered under a 2PL model, the discrimination parameters can. One alternate empirical differential for items under the 2PL model involves the ratio of differences of log odds for two respondents. Specifically, this value:

$$\frac{\text{Log odds}(x_i = 1 \mid \theta_A) - \text{Log odds}(x_i = 1 \mid \theta_B)}{\text{Log odds}(x_j = 1 \mid \theta_A) - \text{Log odds}(x_j = 1 \mid \theta_B)} \tag{1.11}$$

will be constant for any two items i, j, regardless of the proficiencies of respondents A and B (Theorem 21). This means that the ratio between Alice's log odds advantage over Bob



(a) Difficulty parameters in a 2PL model are ordered by the proficiency of respondents with a 50% probability of success on the items.



(b) Under the RP50 order of Figure 1.18a, Item 2 has the lowest difficulty parameter and Item 4 the highest. However, the proficiency region for which respondents have the greatest chance of success on Item 2 (yellow checks) and the region for which they have the lowest chance of success on Item 4 (blue pinstripes) have no overlap. This means that there is no proficiency for which respondents' personal item order matches the RP50 order.

Figure 1.18: Item order in a 2PL model.

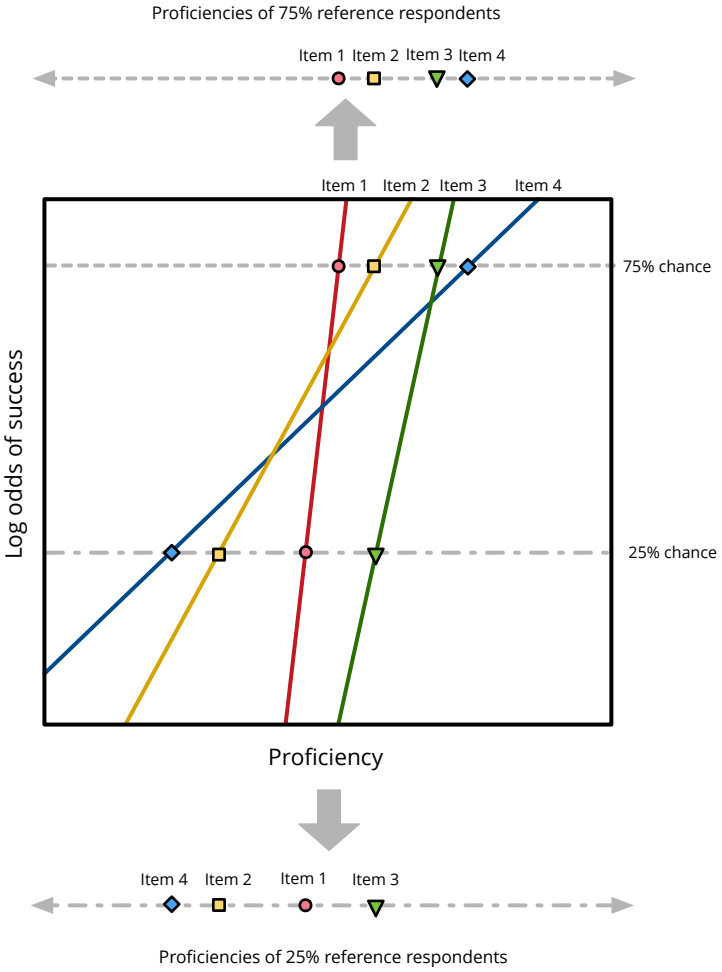


Figure 1.19: Ordering induced on items using different reference respondent populations (RP75 and RP25).

on Item i, and Alice's log odds advantage over Bob on Item j, will be the same if Alice and Bob are replaced by *any other two respondents*, with any other levels of proficiency (Figure 1.20). In either the odds or logit forms, the value of this ratio, given by the expression in (1.11), will be equal to the ratio of the discrimination parameters of the two items (Theorem 20). In the log odds figures, this is equivalent to the ratio of the slopes of the lines.

This value therefore constitutes an empirical, scalar valued empirical differential, and one that is equal to the ratio of the numbers assigned as the discrimination parameters. This means that the discrimination parameters in a 2PL model qualify as a ratio scale under the EDI typology. This classification may be most useful in the field of computer adaptive testing (CAT), where discrimination parameters are often used to aid in item selection (Chang, 2015). Higher discriminating items offer some advantages in distinguishing between close respondents, but are only useful in narrower proficiency ranges, so should not be used too

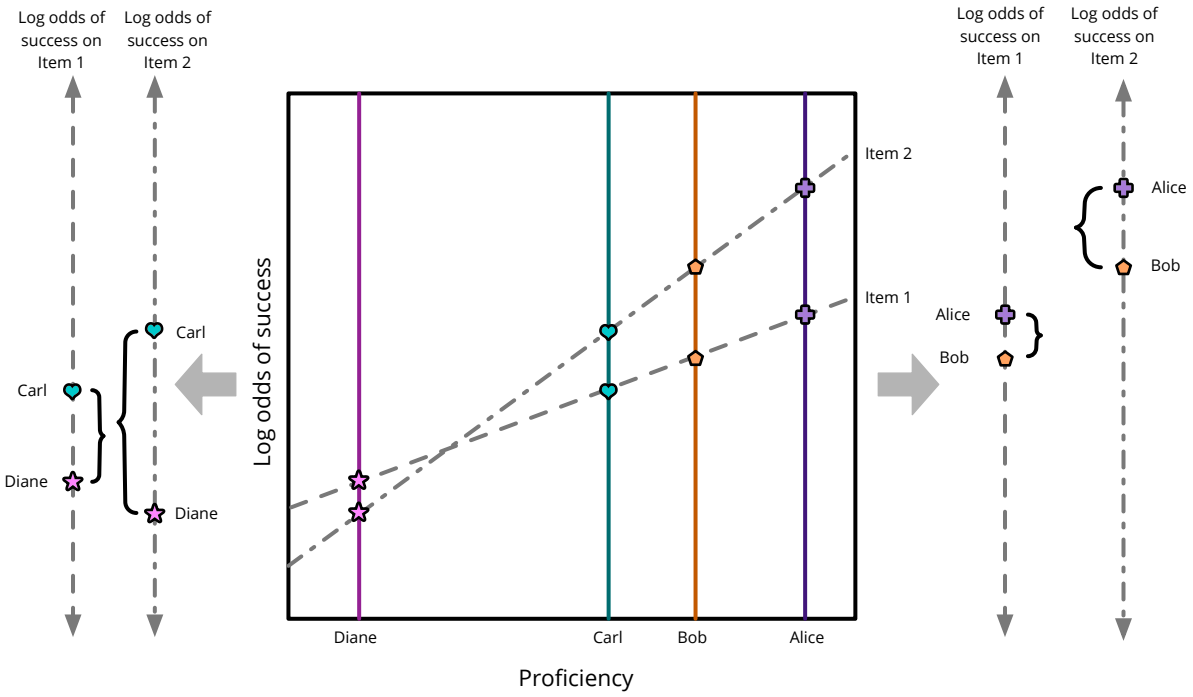


Figure 1.20: The difference between Carl's and Diane's log odds of success is approximately twice as large on Item 2 as on Item 1. This relationship also holds for Alice and Bob. In general in a 2PL model, for any two items, ratios of the differences in log odds between pairs of respondents on the items will be constant, regardless of the pair of respondents chosen.

early in a CAT environment. It is possible that the fact that one item can be meaningfully said to be "twice as discriminating" as another could somehow enable the development of algorithmic item selection methods that rely on these types of calculations. The main implication of this result, however, is that the scale type of 2PL items can be analyzed in multiple ways: as ratio (in terms of discriminations) or as not even nominal (in terms of difficulties).

A better solution for classifying the scale type of items in a 2PL model would take into account the dual structure of these items. Just as lines in the plane have both slopes and intercepts, items have discriminations and difficulty parameters, and are properly understood as binary elements. A multidimensional scale type theory, which would incorporate systems in which objects are assigned not a single number, but ordered pairs or sets of numbers, would be more appropriate. While others have noted the difficulty of applying the Stevens scale types to multidimensional points (Velleman & Wilkinson, 1993), a general framework for these types of scales, that would incorporate structures such as the 2PL item lines, does not appear to have been developed.

1.5 Discussion

Understanding the scale types implied by using statistical models for measurement is crucial to interpreting their estimated results (e.g., measurements of person proficiencies and item properties). When using the log odds form of the Rasch model, these results (when viewed under the ideal slope paradigm) can be interpreted as existing along an absolute difference scale. This makes the value of the subtractive difference between two respondents' estimated proficiency values (e.g. a 1 logit difference) conceivable as an increase in predicted log odds of success for any item in the assessment. As a result, these differences are comparable to those found between estimates of proficiency in other domains, provided those estimates were also obtained using a Rasch model. Estimates from the odds form of the Rasch model can be compared through ratios, which provides potential justification for conclusions such as "Alice has twice the ability as Bob." However, whether these types of statements will be correctly understood by those receiving the results is less clear; people may more intuitively connect this type of relationship to ratios of probabilities rather than odds, but unlike for odds, ratio relationships between probabilities of success do not remain constant as item difficulties vary. Alternately, "twice the ability" may be understood as a ratio of facts known, or speed of response, which are not directly considered by the measurement model.

By contrast, in the 2PL model, the scalar values corresponding to odds ratios or difference in log odds are not in themselves meaningful. However, within a particular set of parameter estimates, these values can be usefully compared. This enables the use of 2PL proficiency estimates in growth measures, impact evaluations, or examinations of the achievement gap. (Note again that *proficiency estimates* under a 2PL model have the interval scale type, although item difficulties do not). Of course, Rasch models proficiency estimates, which are of the ratio scale type, are also suitable for these purposes.

Conversely, the fact that the use of these models engenders such strong conclusions provides practitioners with important considerations when selecting a measurement model: A Rasch model should only be used when they intend to make a strong effort to design, analyze, and modify an instrument until the odds ratio of success between two respondents remains constant across all items. If they do not intend to put in this effort, or expect for theoretical reasons that despite all attempts, this ratio will inevitably vary across different items, then the Rasch model's restriction of a common slope is not appropriate. Similarly, a 2PL model should only be used if practitioners have a theoretical belief that differences in ability should be comparable. Additionally, they should be prepared to undergo a careful and thorough design process aimed at constructing an item set in which comparisons between these differences are stable across items. If this theoretical belief does not apply (for example, if the construct is hypothesized to be only ordinal), or if their design process does not have this invariance as one of its goals, then the parameterization of the IRT models is unsuited in this case.

The associations of models to scale types established in this paper differ from some of those found in the literature. The focus of the next chapter will be on other connections of scale type theory to item response models. This includes an alternate scale typology frame-

work from Suppes and Zinnes (1963), followed by a examination of diverse perspectives on the relationships between scale typology and item response theory that have been described by other researchers, and an analysis of in what manner and for what reasons they differ from those found in this paper.

Chapter 2

Diverse perspectives on scale types in Item Response Theory

2.1 Introduction

The previous chapter presented an argument, based on a formalization of Stevens' original scale typology, that the appropriate scale type classification for the Rasch model is ratio or absolute difference, and the appropriate scale type classification for the 2PL model is interval or relative ratio. The chapter following this one will discuss a framework to unite these respective pairs of scale types, relating them to properties of the attributes being measured.

But before presenting that analysis, I wish to take a step back, to introduce other perspectives on connecting scale types to Item Response Theory. The first of these perspectives, that of a transformation-based typology (Suppes & Zinnes, 1963), will provide a rigorous mathematical foundation for scale type analysis, which I will use both in this chapter to derive an alternate demonstration of the scale type of the IRT models, and in the next chapter to prove properties of the frameworks discussed therein. The second perspective, from Additive Conjoint Measurement (Luce & Tukey, 1964), is the most commonly cited way scale type theory has been connected to IRT in the literature. I discuss the ways in which it has been used to classify IRT scale types, and the reasons why the usual conclusions differ from those of the analyses presented in this dissertation. Finally, I present and discuss an alternate scale type classification of the Rasch model based on the model's derivation from first principles (Fischer, 1995).

This paper is therefore in three, thematically related parts. Its overall goal is to provide an overview of the state of the field regarding connections between IRT and scale type theory, and to attempt to identify, explain, and resolve differences in the conclusions that have been reached on these topics.

2.2 Suppes & Zinnes

2.2.1 Scale type by admissible transformations

In the previous chapter, I described a formalization of the Stevens typology (the "empirical differential isomorphism" or EDI typology) that classified empirical differentials as either scalar values or magnitudes, and then connected those empirical differentials to either subtractive differences or quotients. This formalization has the advantage of maintaining a close connection to Stevens' emphasis on empirical determinations.

However, in applying the EDI typology to psychological measurement, there is an inherent difficulty involved in making empirical observations of what are probabilistic latent traits. In my discussion, I focused on odds of success for respondents and items behaving according to the IRT models, but it is debatable to what degree these odds can be considered "empirical." In this section, I will discuss an alternate formalization of the Stevens scales, developed by Suppes and Zinnes (1963).

Formally, Suppes and Zinnes (1963) define an empirical relational system $\mathfrak{U}$ as a domain of identifiable entities together with a set of relations. For example, a set of weighted objects might have relations indicating order or ratio. This relational system is analogous to Stevens' notion of "empirical properties or relations." In the Suppes and Zinnes (1963) approach, there is again the notion of real-world referents which will be connected to numeric values through measurement.

In Suppes and Zinnes (1963), these numeric values, and their accompanying numeric relations, together comprise the real number relational system $\mathfrak{R}$. A measurement process can then be considered a function f that maps the $\mathfrak{U}$ relational system homomorphically onto the $\mathfrak{R}$ relational system, with the relations defined on $\mathfrak{U}$ mapping to the relations defined on $\mathfrak{R}$. Note that many such functions may be possible for a given $\mathfrak{U}$ and $\mathfrak{R}$. Two weights may be mapped to, say, the values 2 and 4, or alternatively 5 and 10, or any other values that preserve the relational structure. For any two such functions f and g, where f and g are both functions from $\mathfrak{U}$ to $\mathfrak{R}$, we can define $\phi_{fg} : \mathfrak{R} \to \mathfrak{R}$ as the function such that $g = \phi_{fg} \circ f$.

For example, if f represents a measurement function for which the image of an object is its length in feet, and g represents a measurement function that gives length in inches, then for any object $u \in \mathfrak{U}$, we have $g(u) = 12 \cdot f(u)$. The function ϕ_{fg} in this case is given by multiplication by 12 (Figure 2.1). A function h which gave length in centimeters could be related to g through $\phi_{gh}(x) = 2.54x$. In general, all the ϕ functions connecting two length functions will consist of multiplication by a real-valued constant.

This set of all such functions ϕ_{fg}, for all pairs of functions f and g within the set of homomorphic functions from $\mathfrak{U}$ to $\mathfrak{R}$, can be used to define the scale of the attribute being measured. This set is commonly referred to as the "admissible ϕ functions" for the scale.

Suppes and Zinnes (1963) define scale types in terms of the permitted transformations between them, saying that "the type of scale is determined by the relative uniqueness of the numerical assignment." For interval scales, they explain, "instead of asking how we know certain intervals are 'really' equal, we ask if all the admissible numerical assignments are

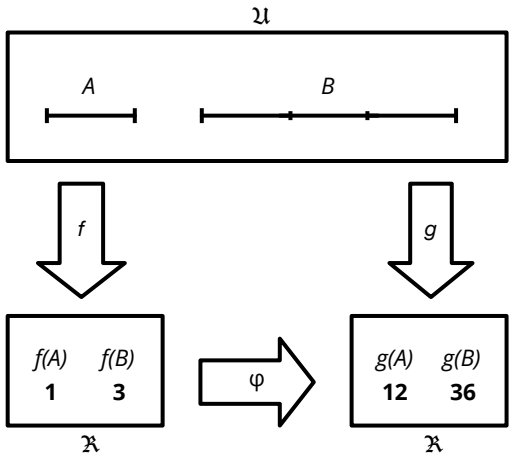


Figure 2.1: Two measurement functions for length, and the transformation between them.

related by a linear transformation." Unlike Stevens, Suppes and Zinnes do not insist that, for example, ordinal scales have a defined order relation, but simply that all ordinal scales defined for a given empirical and numerical relational system be related by a monotone transformation.

These admissible ϕ functions are very similar to what Stevens describes as the transformations preserving the group structure. In most cases, these will be the same sets of functions. A nominal scale is defined under the Suppes and Zinnes (1963) or "SZ" typology as one for which the set of one-to-one transformations specifies the allowable ϕ functions. All the possible measurement functions f and g can be related through ϕ_{fg}, where $g = \phi_{fg} \circ f$, and ϕ_{fg} is any one-to-one transformation. Scales of this type are generally of the "football numbers" variety (as discussed for example in Lord, 1953); the numbers involved simply distinguish individuals or categories, but are not chosen as a function of their magnitudes or mathematical relationships.

Ordinal and nominal scales as defined under the EDI typology will have the same scale types under the SZ system. For ordinal scales, if two measurement maps both have isomorphisms between the $<$ relation and the same empirical order relation, then the transformation between them must be monotone. The converse is not necessarily implied. Since the typology based on ϕ-transformations has no assumption of empirical relations, it will not necessarily be true that for a mapping $f : \mathfrak{U} \to \mathfrak{R}$ within an SZ ordinal scale, $<$ will be isomorphic to some empirical notion of order, even if such a relation exists. However, it is true that *if* such an isomorphism holds for some measurement map $f : \mathfrak{U} \to \mathfrak{R}$, then it will be true for any map in the SZ ordinal scale set (related to f by a monotone transformation).

Suppes and Zinnes (1963) define an interval scale as one for which the set of allowable ϕ functions consist of the set of positive linear transformations. These are the functions of the form $\phi(x) = mx+b$, where m is some positive real number. The temperature transformation $F = \frac{9}{5}C + 32$ between Celsius and Fahrenheit is an example of such a function, as the

temperature scales that lack an absolute zero form an interval scale. As in the ordinal scale case, while an isomorphism between equality of subtraction and an empirical equality of differences is not required by this definition, if one exists for one measurement map f, it will hold for any other map that is related to f by a linear transformation (See Theorem 22, Appendix). However, for interval scales the converse is not necessarily implied: If such an isomorphism exists for two measurement maps, they need not be related by a linear transformation. Strictly speaking, the EDI interval scale type is one that Suppes and Zinnes (1963) label *hyperordinal*: ϕ-transformations preserve order, as well as order of the size of subtractive differences. If certain properties of continuity and density are assumed, an SZ interval scale can be derived. But in a finite case, for example, it need not be present.

The set of similarity transformations (functions of the form $\phi(x) = mx$, where m is some positive real number) defines the SZ ratio scale. Any two ratio scale maps defined as in the previous chapter, in which a real-number valued empirical differential between two elements is equal to the ratio of their assigned values (or a pre-determined function thereof), must be related by such a transformation, so the SZ ratio scale type is implied (Theorem 23). Similarly, any such transformation preserves the value of the ratio between the numbers assigned to any two elements, so ratio scales as defined in the previous chapter and by Suppes and Zinnes (1963) are equivalent structures.

Like Stevens, Suppes and Zinnes do not require any kind of additive relation for ratio scales, explicitly claiming that "fundamental measurement procedures exist that are not based on an addition operation but that lead to ratio (interval or ordinal) scales." By "fundamental measurement procedures," Suppes and Zinnes refer to procedures that do not require any previous measurement. Using a ruler to measure length would be considered fundamental, for example, but deriving density from measures of mass and volume would be considered derived instead. In claiming that fundamental measurement is possible without an addition or concatenation operation, Suppes and Zinnes break from earlier researchers such as Campbell and Jeffreys (1938), who writes that "The properties that are both additive and independent in combination... are the only properties that can be measured fundamentally, that is to say, without previously measuring something else." By contrast, Suppes and Zinnes show that additivity is not one of the required properties for a ratio scale.

Figure 2.2 illustrates the four Stevens scales, and their respective allowable ϕ transformations. This framework for defining scale types opens the door to many other types of scales, beyond the common Stevens typology. The *absolute difference* scales discussed in the previous chapter can now be defined under the SZ typology as scales for which the allowable ϕ transformations consist of functions of the form $\phi(x) = x + b$, with any real number b (Suppes & Zinnes, 1963). Similarly, the *relative ratio* scales can be defined under the SZ typology as scales for which the allowable ϕ transformations are of the form $\phi(x) = b \cdot x^m$.

Overall, the SZ scale typology provides a useful mathematical structure for rigorously defining scale types. These definitions are often easier to apply and to work with than the isomorphism-based system. However, they lose the empirical grounding that connected the scale types to properties of the attributes themselves, making them only dependent on properties of the measurement maps.

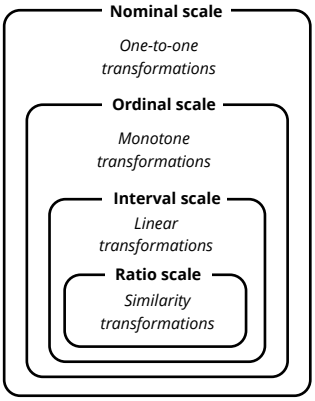


Figure 2.2: The four Stevens scales, and their respective allowable ϕ transformations between measurement maps.

2.2.2 IRT models

In order to connect the Suppes and Zinnes scale typology theory to psychological measurement, it is necessary to identify the following elements within a psychometric framework:

- The empirical relational system $\mathfrak{U}$
- The set of functions $f : \mathfrak{U} \to \mathfrak{R}$

The set of functions ϕ such that $g = \phi_{fg} \circ f$ can then be defined accordingly.

In Item Response Theory, the elements in the empirical relational system can be understood as respondents, or respondent proficiencies.[1] The measurement maps are the assignment functions of proficiency values to respondents.

The Rasch model gives the log odds (logit) of success on an item as:

$$\text{logit}(x_i = 1|\theta) = \theta - \beta_i \tag{2.1}$$

where θ represents a respondent's ability and β_i represents an item's difficulty (Rasch, 1960/ 1980).

Equivalently, the Rasch model can be expressed using odds instead of log odds:

$$\text{Odds}(x_i = 1|t) = \frac{t}{b_i} \tag{2.2}$$

where t and b_i represent the θ and β_i parameters respectively, adjusted for the odds scale. If t and b_i are set to be the exponentiations of θ and β_i respectively, the two models make equivalent predictions.

[1] Arguably, psychometric data should be considered a type of derived measure, rather than fundamental, which would lead to a slightly different setup for the empirical relational system. The scale type analysis is unchanged, however.

The related Two Parameter Logistic (2PL) model adds a so-called item discrimination parameter α_i (Birnbaum, 1968):

$$\text{logit}(x_i = 1|\theta) = \alpha_i(\theta - \beta_i) \tag{2.3}$$

The odds formulation of the 2PL model is:

$$\text{Odds}(x_i = 1|t) = \left(\frac{t}{b_i}\right)^{\alpha_i} \tag{2.4}$$

If the 2PL model is modified such that all the discrimination parameters are constrained to be equal to a constant α_0 across items, the result is a generalized Rasch model (GRM):

$$\text{logit}(x_i = 1|\theta) = \alpha_0(\theta - \beta_i) \tag{2.5}$$

This equation has also been referred to as describing a "family of Rasch models" (Fischer, 1995), in which each choice of discrimination parameter is considered a separate Rasch model. When the value of α_0 parameter is chosen *a priori*, it is referred to as an index rather than a parameter, and the model is a One-Parameter Logistic model (Verhelst & Glas, 1995).

2.2.2.1 Rasch model

In the previous chapter, I identified two possible scale types for the Rasch model, under two different paradigms: The "ideal slope" paradigm, under which it was assumed that all acceptable items measuring the same construct had the same discrimination parameter, and the "selected slope" paradigm, under which items with other discriminations were eliminated or modified for not fitting the model, but were not considered inherently unsuited to measuring the construct in question. This distinction was necessary because of the EDI typology's focus on empirical difference relations. Under the ideal slope paradigm, the constant ratio between two respondents' predicted odds of success under a Rasch model is considered an empirical property of the respondents' ability in the construct, whereas under the selected slope paradigm it is an artifact of a measurement construction process aimed at ensuring such consistency.

Under the SZ typology, the nature of the empirical differential (as a scalar number or a magnitude) is not crucial to scale type classifications. For this reason, the paradigmatic distinction drawn in the previous chapter is less relevant. In this case, the important distinction will be between three different perspectives on Rasch's constant discrimination parameter: as always equal to 1 (in the traditional Rasch model), as a fixed index with a constant value, not necessarily 1 (in the One-Parameter Logistic or 1PL model), or as a parameter that can be changed, as long as it is changed identically for all items in the model (in the Generalized Rasch Model or GRM). This choice will affect which measurement maps are considered valid functions in the scale space.

In fitting an IRT model, there will be a number of sets of estimated parameters, differing only by constraint choices, which can be considered equivalent. In this case, "equivalent"

will mean that they result in identical success probabilities for all persons and items. So, if there are maps $f_\theta, f_\alpha, f_\beta$ which map the empirical person and item attributes to one set of numeric parameters $(\boldsymbol{\theta_f}, \boldsymbol{\alpha_f}, \boldsymbol{\beta_f})$, and maps $g_\theta, g_\alpha, g_\beta$ which map to a different set of numeric parameters $(\boldsymbol{\theta_g}, \boldsymbol{\alpha_g}, \boldsymbol{\beta_g})$, then for these to be considered members of the same scale, both parameter sets must produce equal probabilities of success.

Log odds formulation For the log odds formulation of the Rasch model, for all items i and persons j the following relationship between the parameter sets must hold:

$$\theta_{gj} - \beta_{gi} = \theta_{fj} - \beta_{fi} \tag{2.6a}$$
$$\theta_{gi} = \theta_{fi} + (\beta_{gj} - \beta_{fj}). \tag{2.6b}$$

This implies that if A is an empirical person ability with $g_\theta(A) = \phi_{\theta fg} \circ f_\theta(A)$, then $\phi_{fg}(\theta_f) = \theta_f + c$. Furthermore, the difference between the β values in the two maps is also a constant value across all items. Extending further, for all possible maps f_θ, g_θ and f_β, g_β from attributes to parameter value assignment, the transformations ϕ_θ, ϕ_β from f_θ, f_β to g_θ, g_β will consist of the addition of a numeric constant. Crucially, multiplying by a scale factor will *not* result in equivalent predicted probabilities, so no transformations of that type will be allowed.

Conversely, *any* additive shift of both parameter sets by the same constant c will result in equal predicted probabilities within the model:

$$\theta_{gj} - \beta_{gi} = (\theta_{fj} + c) - (\beta_{fi} + c) \tag{2.7a}$$
$$= \theta_{fi} - \beta_{fj}. \tag{2.7b}$$

Since all valid transformations are additive shifts, and all additive shifts result in valid transformations, the person proficiencies and item difficulties are both SZ absolute difference scales.

These scale types will be the same under the log odds version of the 1PL. In this case, it will be assumed that there is only one allowed map from the constant item discrimination to the reals, with image α_0. Within this scale the following holds:

$$\alpha_0(\theta_{gj} - \beta_{gi}) = \alpha_0(\theta_{fj} - \beta_{fi}) \tag{2.8a}$$
$$\theta_{gj} - \beta_{gi} = \theta_{fj} - \beta_{fi} \tag{2.8b}$$

which is equal to Line 2.6a above, and leads to identical conclusions regarding the restriction of allowable transformations to the addition of a constant. And again, arbitrary transformations of this kind result in valid parameter assignments with no change in the predicted probabilities, provided both sets of parameters (person and item) are shifted identically:

$$\alpha_0(\theta_{gj} - \beta_{gi}) = \alpha_0((\theta_{fj} + c) - (\beta_{fi} + c)) \tag{2.9a}$$
$$= \alpha_0(\theta_{fj} - \beta_{fi}). \tag{2.9b}$$

This means that the log odds form of the 1PL also has the absolute difference scale type. However, under a GRM, different α values are possible for the different maps:

$$\alpha_g(\theta_{gj} - \beta_{gi}) = \alpha_f(\theta_{fj} - \beta_{fi}) \tag{2.10a}$$

$$\theta_{gj} = \frac{\alpha_f}{\alpha_g} \cdot \theta_{fj} + \left(\beta_{gi} - \frac{\alpha_f}{\alpha_g} \cdot \beta_{fi}\right) \tag{2.10b}$$

This is a linear transformation on $\boldsymbol{\theta_f}$. A symmetrical rearrangement shows that a linear transformation has also been applied to $\boldsymbol{\beta_f}$. In fact, any linear transformation $mx+c$, when applied to both the θ and β parameters results in equal predicted probabilities, given one condition: that the reciprocal $\frac{1}{m}$ of the multiplicative factor be applied to the α parameters:

$$\alpha_g(\theta_{gj} - \beta_{gi}) = \frac{1}{m} \cdot \alpha_f((m \cdot \theta_{fj} + c) - (m \cdot \beta_{fi} + c)) \tag{2.11a}$$

$$= \alpha_f(\theta_{fj} - \beta_{fi}) \tag{2.11b}$$

This gives the person and item parameters the interval scale type, when rescaling of the discrimination parameter is allowed within the model.

The EDI typology of the previous chapter also allowed for these two different scale types (absolute difference and interval) for the log odds version of the Rasch model. However, in that case the selection of scale type depended on the paradigm: ideal slope or selected slope. Under the SZ typology, no such paradigmatic assumptions were made. Since scale type depends on the allowable maps and transformations between them, the crucial factor is whether the (common) discrimination parameter is allowed to vary. Allowing this variation removes the possibility of meaningfully describing the difference between respondents in terms of their subtractive difference in proficiency values, since 1 point of difference no longer represents a stable increase in log odds.

Odds formulation If the odds formulation is used, the following relationship holds for the simple Rasch model:

$$\frac{t_{fj}}{b_{fi}} = \frac{t_{gj}}{b_{gi}} \tag{2.12a}$$

$$b_{gi} = b_{fi} \times \frac{t_{gj}}{t_{fj}} \tag{2.12b}$$

In this case, all item parameter values in $\boldsymbol{b_f}$ are transformed to their images in $\boldsymbol{b_g}$ through multiplication by a scalar constant, the value $\frac{t_{gj}}{t_{fj}}$. Likewise, the values in $\boldsymbol{t_f}$ are multiplied by the scalar constant $\frac{b_g}{b_f}$ to become $\boldsymbol{t_g}$. Overall, the set of ϕ transformations between all pairs of scales consist of scalar transformations. Again, we can apply this in the converse to show that all arbitrary scalar transformations result in equivalent model predictions (when applied identically to both parameter sets):

$$\frac{t_{gj}}{b_{gi}} = \frac{m \cdot t_{fj}}{m \cdot b_{fi}} \tag{2.13a}$$

$$= \frac{t_{fj}}{b_{fi}} \tag{2.13b}$$

This formulation of the Rasch model is therefore a ratio scale.

The same applies for the odds version of the 1PL. When the probabilities are assumed to be equal, the transformation must be a scalar multiplication:

$$\left(\frac{t_{fj}}{b_{fi}}\right)^{\alpha_0} = \left(\frac{t_{gj}}{b_{gi}}\right)^{\alpha_0} \tag{2.14a}$$

$$\frac{t_{fj}}{b_{fi}} = \frac{t_{gj}}{b_{gi}} \tag{2.14b}$$

And applying any scalar multiplication results in equal predicted probabilities:

$$\left(\frac{t_{gj}}{b_{gi}}\right)^{\alpha_0} = \left(\frac{m \cdot t_{fj}}{m \cdot b_{fi}}\right)^{\alpha_0} \tag{2.15a}$$

$$= \left(\frac{t_{fj}}{b_{fi}}\right)^{\alpha_0} \tag{2.15b}$$

However, proficiencies and difficulties in a GRM have the relative ratio scale type:

$$\left(\frac{t_{gj}}{b_{gi}}\right)^{\alpha_g} = \left(\frac{t_{fj}}{b_{fi}}\right)^{\alpha_f} \tag{2.16a}$$

$$t_{gj} = \frac{b_{gi}}{b_{fi}^{\left(\frac{\alpha_f}{\alpha_g}\right)}} \cdot t_{fg}^{\left(\frac{\alpha_f}{\alpha_g}\right)} \tag{2.16b}$$

Here, the person proficiencies $\boldsymbol{t_f}$ under the f map are transformed through raising to a power m and multiplication by a constant c. The item difficulties are again symmetrical, so undergo a transformation of the same form. Any transformation of this type preserves the predicted probabilities, with the discrimination parameter multiplied by $\frac{1}{m}$:

$$\left(\frac{t_{gj}}{b_{gi}}\right)^{\alpha_g} = \left(\frac{c \cdot ((t_{fj})^m)}{c \cdot ((b_{fi})^m)}\right)^{\frac{1}{m} \cdot \alpha_f} \tag{2.17a}$$

$$= \left(\frac{t_{fj}}{b_{fi}}\right)^{\alpha_f} \tag{2.17b}$$

This set of transformations defines the relative ratio scale type.

Overall, the scale type analysis shows that under this framework, the scale type of the Rasch model depends on the model formulation. When the model is formulated in terms of log odds, the proficiencies form an absolute difference scale. When it is formulated in

terms of odds, the proficiencies form a ratio scale. Only when the item discrimination is allowed to vary freely between maps, as in the GRM, does the Rasch model yield a mere interval scale. If the item discrimination is fixed at 1, as is commonly practiced, Rasch model parameters cannot be transformed through arbitrary linear transformations without deforming the predicted probabilities as shown:

$$\text{Log odds}(x_i = 1 \mid \theta) = \theta - \beta_i \tag{2.18a}$$
$$\text{Transformed log odds} = m(\theta) + c - (m(\beta_i) + c) \tag{2.18b}$$
$$= m(\theta - \beta_i) \tag{2.18c}$$

If the transformation applied to the item and person parameters includes multiplication by a scalar factor m, the predicted odds of success in the model will change. The model cannot account for this change unless it includes a (universal, in the Rasch case) discrimination parameter which is scaled by $\frac{1}{m}$.

As discussed in the previous chapter, this slope transformation does have interpretational consequences. When the discrimination is held constant at 1, the odds form of the Rasch model ensures that the ratio between two respondents' odds of success on any item is equal to the ratio of their assigned proficiency parameters. This invariance property is a desirable quality for measurement, and echoes measurements of physical quantities such as length or mass, where empirical ratios are similarly expressed through ratios of assigned values. In the log odds form of the model, there is again an invariance represented, this time between difference in log odds of success between two respondents and the subtractive difference between their assigned proficiency parameters. When the discrimination parameter is non-unitary, this invariance is eliminated. This is especially unfortunate under the ideal slope paradigm, where these ratios and differences are assumed to represent a relationship between respondents that is not just relative to the set of items used in the instrument, but is an inherent property of their respective levels of ability in the construct.

2.2.2.2 Two-parameter logistic model

The analysis of Section 2.2.2.1 can be repeated for the Two-Parameter Logistic model. In this model, there are three sets of parameters: person proficiency, item difficulty, and item discrimination. Again, each corresponds to a set of maps from the empirical attributes to numeric proficiencies. Some of these sets of maps will result in identical probability estimates and can be considered equivalent sets. The transformations between the maps in these equivalent sets constitute the ϕ transformations, and will define the scales for proficiency, difficulty, and discrimination.

Log odds formulation For the log odds formulation of the 2PL, let $(\boldsymbol{\theta_f}, \boldsymbol{\alpha_f}, \boldsymbol{\beta_f})$ be the set of numeric parameter values of person ability, item discrimination, and item difficulty under the map set $f_\theta, f_\alpha, f_\beta$, and let $(\boldsymbol{\theta_g}, \boldsymbol{\alpha_g}, \boldsymbol{\beta_g})$ be the set of parameter values under the map set $g_\theta, g_\alpha, g_\beta$. Then the following expresses the relationship between equivalent sets of parameters $(\boldsymbol{\theta_f}, \boldsymbol{\alpha_f}, \boldsymbol{\beta_f})$ and $(\boldsymbol{\theta_g}, \boldsymbol{\alpha_g}, \boldsymbol{\beta_g})$:

$$\alpha_{fi}(\theta_{fj} - \beta_{fi}) = \alpha_{gi}(\theta_{gj} - \beta_{gi}) \tag{2.19a}$$

$$\theta_{fj} = \frac{\alpha_{gi}}{\alpha_{fi}} \times \theta_{gj} + (\beta_{fi} - \frac{\alpha_{gi}}{\alpha_{fi}} \times \beta_{gi}) \tag{2.19b}$$

This means for all persons j, the proficiency parameter θ_{gj} under g_θ is transformed to the f_θ parameter θ_{fj} through multiplication by a scale factor $(\frac{\alpha_{gi}}{\alpha_{fi}})$ and the addition of a constant $(\beta_{fi} - \frac{\alpha_{gi}}{\alpha_{fi}} \times \beta_{gi})$. This transformation constitutes the function $\phi_{\theta fg}$ where $g_\theta = \phi_{\theta fg} \circ f_\theta$ and is a linear transformation. The difficulty transformations ϕ_β are also linear transformations. As in the GRM case, any such linear transformation leaves the predicted probabilities unchanged, if the set of discrimination parameters is scaled to compensate:

$$\alpha_{gj}\left(\theta_{gj} - \beta_{gi}\right) = \frac{1}{m} \cdot \alpha_{fj}\left((m \cdot \theta_{fj} + c) - (m \cdot \beta_{fi} + c)\right) \tag{2.20a}$$

$$= \alpha_{fj}\left(\theta_{fj} - \beta_{fi}\right) \tag{2.20b}$$

The scale for which the ϕ transformations are the set of linear transformations is an interval scale (Suppes & Zinnes, 1963), so for this formulation of the 2PL, the person proficiencies form an interval scale. The fact that the slope can be varied in a 2PL model allows for more freedom in transformations than in the Rasch model case where the slope is held constant at 1.

The discrimination transformations ϕ_α between different maps f_α consist of scalar multiplications, making the f_α maps a ratio scale.

Odds formulation The odds formulation of the 2PL model gives the following relationships:

$$\left(\frac{t_{fj}}{b_{fi}}\right)^{a_{fi}} = \left(\frac{t_{gj}}{b_{gi}}\right)^{a_{gi}} \tag{2.21a}$$

$$t_{fj} = b_{fi}(b_{gi})^{-\frac{a_{gi}}{a_{fi}}} \times \left((t_{gj})^{\left(\frac{a_{gi}}{a_{fi}}\right)}\right) \tag{2.21b}$$

In this case, the transformation from t_{gj} to t_{fj} involves exponentiation to the power $\frac{a_{gi}}{a_{fi}}$ and then multiplication by the scale factor $b_{fi}(b_{gi})^{-\frac{a_{gi}}{a_{fi}}}$. The difficulty parameters are themselves exponentiated and multiplied by a constant in their ϕ_b transformation, and any such transformation preserves the predicted probabilities, when an appropriate scalar multiplication is applied to the set of discrimination parameters:

$$\left(\frac{t_{gj}}{b_{gi}}\right)^{\alpha_{gi}} = \left(\frac{c \cdot ((t_{fj})^m)}{c \cdot ((b_{fi})^m)}\right)^{\frac{1}{m} \cdot \alpha_{fi}} \tag{2.22a}$$

$$= \left(\frac{t_{fj}}{b_{fi}}\right)^{\alpha_{fi}} \tag{2.22b}$$

	Log odds formulation	Odds formulation
Rasch model	Absolute difference scale	Ratio scale
GRM	Interval scale	Relative ratio scale
2PL model	Interval scale	Relative ratio scale

Table 2.1: Proficiency scales

This is then also a relative ratio scale. Table 2.1 shows the relationship with the way person proficiencies are represented in various models.

Thus, just as there are two possible scales for the Rasch model, depending on the formulation, there are also two possible scales for the 2PL model: interval and relative ratio. Each pair of models is equivalent, with parameters in one member of the pair equal to the log of the parameters in the other member. Under the SZ framework, these are each considered separate scales. In the next chapter, I will look at a framework that connects these type of isomorphic scale types.

2.2.3 Practical example

One example case in which scale type analysis could be helpful is in the Delta-Dimensional Alignment (DDA) method, designed for increasing comparability between parameters in different dimensions (Schwartz & Ayers, 2011). In the first step of this technique, a multidimensional dataset is analyzed using a single unidimensional Rasch model, as if all items were in the same dimension and were using the same construct. From this model, the mean ($\mu_{d(uni)}$) and standard deviation ($\sigma_{d(uni)}$) of the items in each dimension d are calculated and retained.

Following this analysis, another model is run on the same dataset. This model is run as a standard Multidimensional Random Coefficients Multinomial Logit (MRCML) model (Adams, Wilson, & Wang, 1997), with the items assigned to the appropriate dimensions. Considered separately, each dimension is a Rasch model. Within each dimension, the mean item parameter is constrained to be zero. The estimated difficulty parameters are then labeled $\beta_{i(multi)}$. The standard deviation of the item parameters for each dimension is calculated as $\sigma_{d(multi)}$.

In the next step, the item difficulty parameters from the multidimensional model are adjusted based on the properties of the unidimensional model parameters. Specifically, each item difficulty parameter β_i for an item in dimension d is modified in the following way:

$$\beta_{i(DDA)} = \beta_{i(multi)} \times \frac{\sigma_{d(uni)}}{\sigma_{d(multi)}} + \mu_{d(uni)} \tag{2.23}$$

The final step of the DDA process is to re-estimate the person parameters, using the anchored item parameters calculated in the previous step. This estimation again uses a standard MRCML model.

The re-scaling performed in Equation 2.23 is a linear transformation. However, there is no corresponding transformation of the discrimination parameter. This means that the resulting predicted probabilities of success will be different. Since the goal of DDA is simply to make parameters comparable, and has no theoretical reason to support changing the model predictions entirely, this is an undesirable side effect. It is also one that has often been ignored in applications of DDA.

In the context of a science argumentation instrument, researchers described using DDA "to calibrate the scientific argumentation and general argumentation dimensions onto a common metric" (Osborne et al., 2016) and claimed that "The metrics of these two dimensions have been transformed into a common scale" (Yao, Wilson, Henderson, & Osborne, 2015). Similarly, researchers analyzing a teacher licensure exam reported using DDA "to place items from all three dimensions on the same scale" (Castellano, Duckor, Wihardini, Telléz, & Wilson, 2016). Similar reasoning was used by researchers in regards to instruments measuring understanding of the structure of matter (Morell, Collier, Black, & Wilson, 2017), attitudes toward religion (Hermisson, Gochyyev, & Wilson, 2018), and acceptance of evolution (Sbeglia & Nehm, 2019). The implication seems to be that rescaling model parameters in this way is analogous to rescaling parameters in a normal regression model; that the model will simply adjust appropriately to compensate and the results will be equivalent mathematically, but with more desirable properties in terms of interpretation. However, the fact that the discrimination parameter is not allowed to shift in the DDA process means that the model cannot adjust to these transformations, and is negatively distorted.

One solution is to modify the DDA process by allowing a non-unitary discrimination parameter. This variant process, used by (Feuerstahler & Wilson, 2019), adds a discrimination parameter to the re-scaled model, defined as follows:

$$\alpha_{d(SADDA)} = \frac{\sigma_{d(multi)}}{\sigma_{d(uni)}} \tag{2.24}$$

Essentially, this variant addresses the issues with the DDA method by performing a complete re-scaling. While this adjustment is newly suggested and has yet to catch on, it is possible that a better understanding of the Rasch model scale type, and its associated allowable transformations, would hasten its acceptance.

The associations of models to scale types established in this and the previous chapter differ from some of those found in the literature. The next section will focus on diverse perspectives on the relationships between scale typology and item response theory that have been described by other researchers, and an analysis of how and why they differ from those found in this section and the preceding chapter.

2.3 Additive Conjoint Measurement

The Stevens (1946) system is well suited to cases where the empirical relations between attributes and objects are observable. In psychological measurement, when all we have

to observe are responses to items, such relationships may not be as clear. The simplest observation which can often be made is that, in general, over multiple item sets ostensibly measuring the same construct, one respondent will tend to respond correctly more often than another. Similarly, we may observe that one item is solved correctly more often than another when presented to multiple different groups of respondents. If the only observable comparisons are ordinal, it seems reasonable to suppose that the only underlying group structure that can be deduced is itself ordinal. In fact, it is in some cases possible to derive an interval structure using only ordinal relationships. Additive conjoint measurement (Luce & Tukey, 1964) describes a structure which permits such a derivation.

Additive conjoint measurement applies in a very specific case in which the observed data are themselves combinations of two elements. So, an observed element (A, P) is a combination of $A \in \mathscr{A}$ and $P \in \mathscr{P}$. The classic example is density, which is a combination of mass and volume. Later, I will discuss examples from psychometrics, in which responses to items can be seen as combinations of items and persons.

These binary elements have an ordering relation $\geq$ that is reflexive, transitive, connected, and antisymmetric. This means that for all $A, B, C \in \mathscr{A}, P, Q, R \in \mathscr{P}$

1. $(A, P) \geq (B, Q)$ and $(B, Q) \geq (C, R)$ imply $(A, P) \geq (C, R)$;
2. Either $(A, P) \geq (B, Q)$ or $(B, Q) \geq (A, P)$ or both;
3. $(A, P) = (B, Q)$ if and only if $(A, P) \geq (B, Q)$ and $(B, Q) \geq (A, P)$.

In addition to order, the observed elements have a property often referred to as *solvability* (Narens & Luce, 1986). This property says that if any three of the four elements A, B, P, Q are specified with $A, B \in \mathscr{A}$ and $P, Q \in \mathscr{P}$, the equation $(A, P) = (B, Q)$ can be solved for the fourth.

The third axiom, referred to as *double cancellation*, provides a way to think about comparing intervals. Essentially, while we cannot compare intervals AB and CD directly (with $A, B, C, D \in \mathscr{A}$) we can compare them to intervals PQ with $P, Q \in \mathscr{P}$. Suppose we have a paired element (A, Q). We can imagine changing it in two ways: By changing A to B, or by changing Q to P. If the increase caused by changing from Q to P is greater than the increase from A to B (or, equivalently, if the decrease is smaller) then we will have $(A, P) \geq (B, Q)$. We can thus interpret $(A, P) \geq (B, Q)$ to mean that the interval QP is greater than AB.

If we follow this interpretation to take the statements $(A, X) \geq (F, Q)$ and $(F, P) \geq (B, X)$ to mean that $QX \geq AF$ and $XP \geq FB$ respectively, then the double cancellation axiom, which says that in the above case we must also have $(A, P) \geq (B, Q)$, implies that these intervals can be effectively concatenated to yield $QP \geq AP$. Essentially, while we have only made ordinal observations, we have arrived at an interval scale structure.[2] Crucially, this structure applies not only to the binary (A, P) elements, but to each factor set as well.

[2]Another axiom, the Archimedean property, is necessary to complete the structure, but is outside the scope of this paper. It also requires only ordinal observations on the (A, P) elements.

Overall, the theory of conjoint measurement states that when these axioms hold, there exist real-valued functions T, f, g such that for all $A \in \mathscr{A}, P \in \mathscr{P}$:

$$T(A, P) = f(A) + g(P) \tag{2.25}$$

and that all possible such functions are related by linear transformations.

2.3.1 Connecting ACM to Rasch

A response to an item can be thought of as a combination of two factors: the item, and the respondent. For this reason, the theory of Additive Conjoint Measurement has a natural application to psychological measurement. There are two unobserved factors, and an observed combination (the response). While this structure describes many item response models, the log odds form of the Rasch model, which hypothesizes item response probabilities as an additive combination of the item and person parameters, is particularly similar. As Michell (1986) puts it, "the sort of situation described by the Rasch model is an instance of the sort of situation treated by the theory of conjoint measurement."

In fact, the connection between ACM and Rasch is quite strong. If $\mathscr{A}$ is considered to represent person proficiencies, and $\mathscr{P}$ to represent difficulties, then paired elements (A, P) with $A \in \mathscr{A}$ and $P \in \mathscr{P}$, can be considered to represent the probability that a respondent with proficiency A will succeed on an item with difficulty P. Defined in this way, the estimated response probabilities produced by fitting a Rasch model will always satisfy the order restrictions specified by the conjoint additivity axioms (Karabatsos, 2001). As real numbers, estimated response probabilities have a natural order relation ($<$) which is reflexive, transitive, connected, and antisymmetric. It can also be shown that the predicted probabilities obey the double cancellation axiom (Theorem 24). As for solvability, there will always be possible levels of proficiency and difficulty that will solve any given equation, although there may not exist respondents or items with these values in any given dataset. Together, this means that the ACM axioms apply to the system of probabilities predicted by the Rasch model.

For this reason, the Rasch model can be seen as a possible solution to one of the problems of ACM: finding the appropriate interval scaled mapping function T from the combined elements (A, P) to the real numbers. The Rasch model suggests using log odds of success. This is how Keats (1967) presents it:

> [A] function [meeting the constraints for conjoint measurement] must be found.... A function which is intuitively attractive... has been proposed by Rasch."

This connection between Rasch and ACM, and the further connection between ACM and interval scales, can be used to establish the Rasch model as interval scale measurement, with reasoning such as the following:

> For most properties that are of interest to psychology, interval level measurement is unattainable. There are circumstances however where such measurement is possibly within reach. These circumstances are captured by the assumptions of the well-known Rasch Model. This model, according to many, is an instantiation of an additive conjoint representational measurement structure. This is a structure for which it was shown (Luce & Tukey, 1964) that if its axioms hold, the representation is of the interval level. (Zand Scholten, 2011)

There are numerous other examples in the literature referring to the Rasch model as an example of conjoint measurement (see Kyngdon, 2008, for an extensive list). However, it is not precisely clear what the implications of this connection may be. While the predicted probabilities of the Rasch model comply to the ACM axioms, the mere act of fitting a Rasch model to data does not guarantee that there are underlying constructs that behave as desired. One possible approach is suggested by Brogden (1977), who sees evaluating the fit of a Rasch model as an alternative to trying to prove the ACM axioms. He claims "It is reasonably obvious that a fit of the Rasch model implies that the cancellation axiom will be satisfied.... It then follows that items and persons are measured on an interval scale with a common unit." He is not alone in this belief; other researchers describe "testing how well test data fit the Rasch model, and hence satisfy the requirements of probabilistic additive conjoint measurement" (Preece, 2002).

The reasoning here seems to be that if the Rasch model shows good fit according to relevant fit statistics, then it must be "true," or at least close enough to true that the "real" probabilities will surely behave like the predicted probabilities and must therefore conform to the ACM axioms. This reasoning is made explicit in Vessonen (2018), who outlines the following postulates, leading to a conclusion that tests of fit with the Rasch model are evidence of interval scale representation:

P1 If we have evidence that the axioms of conjoint measurement are fulfilled then we have evidence of an interval level representation of the attributes of interest.

P2 If we have evidence that the attribute of interest has the structure postulated in the Rasch model, then we have evidence that manifestations of the attributes fulfill the axioms of conjoint measurement.

P3 If empirical tests of fit between data and the Rasch model show that the data fits the model, then we have evidence that the attributes of interest have the structure postulated in the Rasch model.

A similar line of reasoning can take model fit as a direct test of whether an attribute is quantitative:

> Nobody working in IRT, and we dare to make this statement as a universal claim, accepts the hypothesis that attributes are quantitative without testing the model

> for its empirical adequacy. As a matter of fact, IRT models are regularly rejected because they do not adequately fit the data. (Borsboom & Mellenbergh, 2004)

In fact, models can show good fit even for data for which the Rasch model is known to be inappropriate. Karabatsos (2001) uses work by Nickerson and McClelland (1984) to conclude that "it is possible for a numerical conjoint measurement model, such as the Rasch model, to conclude excellent or perfect data fit, even for data sets containing serious violations of the conjoint measurement axioms."

The Rasch model may also fit even when the initial assumption, that there are underlying quantitative attributes that combine to determine response probabilities, is unwarranted. Michell (2004) claims that "these models may fit even when the relevant attributes are non-quantitative." Michell's definition of "quantitative" includes only interval or ratio scales, and specifically excludes randomly generated data (e.g., coin flips as in Wood, 1978) as well as ordinal data.[3]

For this reason, Michell (1997, and others) views the task of establishing an attribute as quantitative to be the essential first step of measurement, since "Until the scientific task of quantification is completed, claiming that a procedure measures anything is premature." Under this point of view, the connection between the ACM axioms and the Rasch model is reversed. Instead of seeing a well-fitting Rasch model as a sign that the ACM axioms apply (and the data is therefore quantitative), the compliance of the data to the ACM axioms is used to establish the data as quantitative (Michell, 2000). Once established as quantitative, an attempt can be made to apply the Rasch model (Michell, 2008). Then the double cancellation axiom, for example, constitutes a prediction such that "If this prediction is confirmed, then this supports the hypothesis that [the attributes] are quantitative; if infirmed, then not" (Michell, 2000).

One complication is that the task of empirically testing the ACM axioms is made difficult when the traits in question are latent and probabilistic. As Perline et al. (1979) note, "obviously, the p_{ij} [probabilities] are unobservable, as are the ability and item parameters." This represents a clear difference between the Rasch model and the structure described by the ACM axioms: None of the three sets of properties (proficiencies, difficulties, or probabilities) are directly observable. To address this limitation, Perline et al. (1979) use properties of the Rasch model to outline an empirical strategy towards verifying the ACM axioms. Because total scores constitute a sufficient statistic of respondent ability, they begin by treating all respondents with the same raw score as if they have the same latent ability level. They can thus use the success rates within a group of equivalent respondents as approximations of the true probabilities, and use these proportions to test the ACM axioms. For monotonicity,

[3]Michell (2004)'s example of ordinal data is Bond and Fox (2001), which he considers ordinal due to its explicit grounding in Piaget's stages of development. However, quantitative models of these stages are also possible (Draney, 1996), so this is not an ideal counterexample for Michell's purposes. It is also not clear that the kind of randomly generated data described in Wood (1978) is a good example of a non-quantitative attribute, either.

they use Kendall's coefficient of concordance to test for rank agreement between pairs of rows and columns. For cancellation, they computed the rate of violations per 3×3 submatrix.

The Perline et al. (1979) method has its drawbacks. As the authors note, the grouping approach "relies on statistical theory, not conjoint measurement theory, for justification." Furthermore, since the observed proportions are only estimates of the underlying probabilities, violations of the ACM axioms are likely to occur even in the case where the additivity hypothesis is true, and it is not clear what level of violations should be considered acceptable.

A more comprehensive approach to axiom validation was developed by Karabatsos (2001), based on the Scheiblechner (1999) probabilistic analogue of the ACM axioms. In this method, the Markov Chain Monte Carlo method is used to generate a distribution space for the matrix of respondent-item response probabilities, with ordinal restrictions applied corresponding to the ACM axiom constraints. Observed proportions that fall outside the middle 95% of the generated distribution are taken as indications of axiom violation. An improved sampling algorithm was developed by Domingue (2014), with a variant using synthetic likelihood proposed by Karabatsos (2018).

2.3.2 Applications to the 2PL

As discussed above, the predicted probabilities of the Rasch model conform to the axioms of additive conjoint measurement. However, the 2PL model does not behave in the same way. Specifically, when ordering items in terms of predicted probability of success, two respondents of different abilities may have different item orders. This contradicts the order axioms of ACM.

This contradiction has been taken as evidence that the 2PL cannot have interval scale properties. Cliff (1989), for example, writes that the authors of ACM "provide axioms that define necessary and sufficient conditions for an interval scale." He concludes that in the 2PL "Consistency of orders is violated so double cancellation is violated so the necessary condition for an interval scale is violated."

While it is true that the lack of order consistency prevents the 2PL from conforming easily to the axioms of ACM, the stronger conclusion, that the proficiencies are not interval, does not follow. First of all, while the ACM axioms provide *sufficient* conditions for interval scale measurement, Cliff is incorrect to state that Luce and Tukey (1964) show that they are *necessary* conditions. Second of all, modified procedures can reconcile the 2PL model with the ACM axioms. For example, items within the model can be stratified by their discrimination parameters. Within each stratified set, the pairs of proficiency and discrimination parameters will behave just as within a Rasch model, thereby inducing, per the reasoning discussed above, interval scales on both sets. For the proficiency parameters, the interval scale induced by the pairing with each stratified set of item parameters will be consistent, with equality of order and equality of differences remaining constant. For the item parameters, it will not be possible to unify the different stratified sets into a single interval scale, but this does not make the proficiency parameters any less interval. It is true that this strategy involves some unlikely assumptions: It is rare for an item set in the 2PL model to contain *any* items of

equal discriminations, let alone suitably large classes of them to support application of the ACM axioms. However, the original ACM axioms contain an equally impossible existence postulate: the solvability axiom. If items and respondents at the appropriate levels can be asumed to exist in theory to satisfy that axiom, then additional items with the necessary discriminations can also be presumed.

Alternately, additional axiomatic structures can be used for the 2PL. An extension to ACM, known as Polynomial Conjoint Measurement or PCM, incorporates additional parameters (Tversky, 1967). If D is a subset of $(A \times B \times \ldots K)$, where $A, B, \ldots K$ are a finite number of disjoint sets, and there is a binary relation on D which establishes a partial order, then D satisfies a polynomial measurement model M if there exists an order-preserving real-valued function f on D and real-valued functions $f_A, f_B, \ldots f_K$ on $A, B, \ldots K$ respectively such that for any $(a, b, \ldots k) \in D$, $f(a, b, \ldots k) = M(f_A(a), f_B(b), \ldots f_K(k))$, where M is a polynomial function. For three parameters, possible simple polynomials include $M = f_A(a) + f_B(b) + f_C(c)$ (additive), $M = f_A(a)(f_B(b) + f_C(c))$ (distributive), $M = f_A(a)f_B(b) + f_C(c)$ (dual-distributive), or $M = f_A(a)f_B(b)f_C(c)$ (multiplicative) (Krantz & Tversky, 1971).

The distributive (or under some parameterizations, dual-distributive) model has the same structure as the 2PL. Ballou (2009) used this to connect polynomial conjoint measurement to the 2PL model, claiming "The empirical relations between examinees and items are termed a *polynomial conjoint structure*" (emphasis original), and noting that under it the item and difficulties and person parameters would have interval scale properties. Kyngdon (2011) illustrated a possible procedure for validating the empirical structure in the PCM case.

2.3.3 Discussion

The application of conjoint measurement theory to psychometric models has provided several promising avenues for assumption checking. The framework developed by Karabatsos and others is able to identify structural violations beyond the capabilities of the usual model fit parameters.

However, it would be a mistake to conclude that because the Rasch model can be taken to comply with the conjoint measurement axioms and the 2PL model does not, that the Rasch model is therefore an example of interval measurement and the 2PL model cannot be interval. While there are certainly parallels between conjoint measurement and IRT models, the analogy is not perfect, even in the probabilistic extension.

In the first approach discussed above, in which good fit of the Rasch model is taken as evidence for satisfaction of the ACM axioms and hence interval scale measurement, the medial step is unnecessary. To borrow the postulate structure of Vessonen (2018), Postulates 1 and 2 could be combined to yield "If we have evidence that the attribute of interest has the structure postulated in the Rasch model, then we have evidence of an interval level representation of the attributes of interest" without reference to ACM at all. This was shown in the discussion in Section 2.2.2.1, as well as the previous chapter, which in fact both proposed a stronger result: That given certain paradigms or model constraints, an

attribute with the structure postulated in the Rasch model had ratio or absolute difference level representation. The detour to ACM merely weakens the conclusion of this argument.

In the second, reversed approach, in which satisfaction of the ACM axioms is used as evidence of quantitivity, thereby justifying employment of the Rasch model, it should be noted that conjoint measurement is designed for the case in which the only observations that can be made are ordinal. There are certainly reasons why ordinal relations can be more readily considered observable in a psychometric context than other relationships between levels of respondent ability. When using a Rasch model, the numerical estimates of respondent ability will have the same order as respondents' observed summed item scores. For this reason, it seems reasonable to treat respondent order as "known." In contrast, success frequencies may be considered less "observable" in this way, since it will generally not be possible, in practice, to assign numerical values to proficiency and difficulty levels in such a way that the predicted success frequencies will be equal to the observed success rates for various sets of respondents. Instead, estimated values can be chosen to maximize the likelihood of the observed frequencies, or to be the most probable values held by respondents based on some prior expected distribution. These decisions affect the success frequencies predicted by the model, and in turn many of the relationships between respondents (e.g., "twice the odds of success"), but leave ordinal relationships unchanged. These facts complicate the feasibility of using observed success frequencies, and the relationships between them, as a proxy for the latent "true" frequencies, in the same way that the observed order of sum scores is taken as an indication of the true order of respondent abilities.

However, this true order, like all other aspects of the latent underlying probabilistic attributes, is not directly observable in the manner expected by the ACM axioms. The fact that the observed order matches the order of the estimated values is nice, and certainly adds to the face validity of the results, but it does not mean this true order has been observed. Values such as "the odds ratio between two respondents that makes their observed success frequencies the most likely" require more extensive procedures to compute, but are arguably as much based on observables in the data as is the respondent order. For these reasons, I believe that the line implied by the use of ACM that places "order" on the observable side and everything else as unobservable is not necessarily a valid distinction to make in considering psychometric data.

Additionally, as noted above, the scale type of the Rasch model is ratio, rather than interval. This means that even if we permit the use of ACM axiom verification on response frequencies to establish interval scale structure, the use of the Rasch model is not necessarily justified by positive results. An interval scale structure is weaker than that required by the Rasch model equation.

Finally, it is important to remember that the failure of the 2PL model to comply with these axioms is not because proficiencies are not measured along an interval scale in this model. It is because items are not located on an interval scale. Items in a 2PL model have both difficulty and discrimination. Strictly speaking, the 2PL model does not fully conform to the axioms of polynomial conjoint measurement, either. In PCM, each set of attributes are able to be varied independently and ordered accordingly. In a 2PL, items of

different discriminations cannot be meaningfully ordered according to difficulty. The usual model used for the 2PL orders item difficulties according to the level of ability required to have a 50% chance of success on the item. However, choosing 30% or 70% as the threshold would result in a different item order, in which higher discriminating items become harder or easier, respectively, relative to lower discriminating items. This makes it difficult to apply the axioms of PCM, which require unambiguous ordering, to *items* in 2PL. However, none of this prevents 2PL *proficiencies* from having interval scale qualities.

2.4 Derivations from special objectivity

In addition to using the theory of Additive Conjoint Measurement, other connections have been drawn between scale type theory and Additive Conjoint Measurement. One is from Fischer (1995), who derives the Rasch model from several properties and concludes that "scales for person and item measurement are *interval scales*" (emphasis original). In Fischer's case, his derivations lead, as he writes, "to a 'family of RMs,' that is, to a logistic model where all items have the same (unspecified) discrimination parameter." In other words, Fischer (1995) has derived the Generalized Rasch Model given in Equation 2.5.

The use of non-unitary discrimination parameters in the GRM does allow for more generalized linear transformations, as discussed in Section 2.2.2.1, resulting in the interval scale type. As mentioned in that section, this type of transformation is reasonable if the constant odds ratio between respondents is seen as an artifact of item selection, rather than an inherent property of the respondents (as in the selected slope paradigm). Otherwise, there is no reason to use a discrimination parameter other than 1, any more than it is useful to use a ruler in which all the values at the tick marks have been squared.

Fischer (1995)'s argument against the ratio/absolute difference scale claim is as follows:

> The measurement properties must be determined by empirically testable laws, such as SO [specific objectivity]. Whether SO is introduced as a postulate, or whether it follows from the assumption of sufficiency of the raw score, is immaterial, because in both cases it is an empirically testable model property.... Such tests, however, are not sensitive to changes of the common discrimination parameter of all items and hence cannot be used as arguments for specifying [discrimination] = 1.

Specific objectivity is the notion that the result of a comparison between two objects should not depend on the elements used for comparison (Rasch, 1966). When applied to the Rasch model, the principle of specific objectivity means that a comparison between two respondents' ability levels should not depend on the items administered to them, and a comparison between two items' difficulties should not depend on the respondents who completed them.

The Fischer (1995) derivation of the GRM from specific objectivity starts from the assumption that a comparison can be made between two respondents' response probabilities

on a given item, and that the value of this comparison function is independent of the item chosen. This assumption, together with a few other structural assumptions, lead him to derive for the Item Characteristic Curve (ICC) the following form (Equation 2.63 in the original; notation slightly adjusted):

$$P(x_i = 1 \mid \theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \tag{2.26}$$

This is simply the probabilistic form of the logistic Rasch model, as can be seen when it is rearranged:

$$\begin{aligned}
\text{Log odds}(x_i = 1 \mid \theta) &= \log\left(\frac{P(x_i = 1 \mid \theta)}{1 - P(x_i = 1 \mid \theta)}\right) && \text{(2.27a)}\\
&= \log\left(\frac{\left(\frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)}\right)}{1 - \left(\frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)}\right)}\right) && \text{(2.27b)}\\
&= \log\left(\frac{\left(\frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)}\right)}{\left(\frac{1}{1+\exp(\theta-\beta_i)}\right)}\right) && \text{(2.27c)}\\
&= \log(\exp(\theta - \beta_i)) && \text{(2.27d)}\\
&= \theta - \beta_i && \text{(2.27e)}
\end{aligned}$$

Fischer claims that the representation is unique "up to linear transformations $a\theta + b_1$ and $a\beta + b_2$." Strictly speaking, this does not follow. While θ and β can be shifted by a (common) additive constant, any multiplication will deform the ICC, as will shifts by non-equal b_1, b_2. This is simple to see in the log odds form, but can also be derived from the probabilistic form. Fischer cites an earlier lemma in the paper (his 2.2) for the linear transformation claim, in which he shows that if a model has ICCs of the form $f(\theta - \beta)$, then linear transformations can be applied to the θ and β values. The difference is that in this lemma, Fischer allows the transformed values to have a different ICC function and gives the general form (Equation 2.22 in the original):

$$P(x_i = 1 \mid \theta) = \frac{\exp[a(\theta - \beta_i) + b]}{1 + \exp[a(\theta - \beta_i) + b]} \tag{2.28}$$

These constants a and b are conspicuously missing from the derivation from specific objectivity. Their absence is due not to any properties of SO itself, but to two simplifying assumptions Fischer makes along the way in his derivation: First, that the θ values can be defined by choosing a reference item, and setting each respondent's proficiency value as the log of their odds of success on that item. Second, that the β values can be defined by the difference in a respondent's log odds of success on that item, and on the reference item. (The fact that this difference value does not depend on the choice of respondent is derived as a

consequence of the specific objectivity assumption.) The first of these simplifying assumptions, together with the SO-derived fact that differences in log odds between respondents are constant across items, eliminates the possibility of multiplying the parameters by some non-unitary scalar factor and thereby restricts the parameter sets to absolute difference scales. The second assumption eliminates the possibility of using different b_1, b_2 additive shifts for the θ and β values, but does not affect their scale type. Neither assumption is strictly necessary for Fischer's derivations, which is perhaps why he concludes that arbitrary linear transformations are permitted by the derivation from SO.

Incidentally, other derivations from specific objectivity have concluded that the Rasch model has ratio scale properties. Wright and Linacre (1987) made equivalent assumptions to those in Fischer (1995) using the odds form of the Rasch model:

1. Choose a reference person p_0 and a reference item i_0, such that p_0 has even odds (50% chance) on item i_0.

2. For every person, let their proficiency parameter be their odds on item i_0.

3. For every item, let its difficulty parameter be the odds of person p_0 on that item.

This formation results in a ratio scale. For it to be a relative ratio scale (the exponentiated equivalent of a ratio scale), steps 2 and 3 would have to involve raising the odds value to an arbitrary power (the discrimination parameter).

At any rate, Fischer's objection, quoted above, is that assumptions such as SO are testable, while it is not possible to test whether the common discrimination parameter is equal to 1 or some other value. This objection is perhaps not very well stated. The value of the common discrimination parameter has no effect on the model predictions, provided the other parameters are scaled accordingly. In that sense it is of course not "testable" in the same way it is not "testable" which, if any, of the difficulty parameters is set to the value 0. It is a choice about parameterization, rather than an assumption or observation about the underlying data. Perhaps Fischer means, then, that like the choice of which item to use as a reference, the choice of discrimination parameter is arbitrary. Just as choosing a different reference item shifts the parameters by an additive constant, the choice of a different discrimination parameter scales the parameters by a multiplicative factor. In the odds form of the model, it raises the parameters to a given power. If 1 is seen as an arbitrary value, then the parameters are on interval scales in the log odds case, and relative ratio scales in the odds case.

The flaw in this reasoning is that 1 is not, in fact, an arbitrary value. The choice of 1 is necessary to preserve a simple relationship: That the difference in log odds between two respondents be equal to the difference in their proficiency parameters. In the odds form of the model, this relationship is that the ratio between respondents' proficiency parameters is equal to their odds ratio of success. This is the same kind of relationship observed in other ratio scales, such as length or mass. It is mathematically possible to transform these parameters such that this relationship does not hold, provided a corresponding transformation is made

through the discrimination parameters. But by the same reasoning, it would be possible to square every number on a yardstick, provided a "discrimination" exponent of $\frac{1}{2}$ was applied when needed, yielding new formulas such as:

$$\text{Slope} = \left(\frac{y}{x}\right)^{\frac{1}{2}} \tag{2.29a}$$

$$\text{Circumference} = \pi^2 \times d \tag{2.29b}$$

This measurement system is perfectly consistent, and there is no way to "test" that the length values it produces are incorrect. But numbers on a ruler are never transformed in this way, both because it is cumbersome, and because it obscures the relationships between objects that the measurement process is intended to elucidate. The same is true for a Rasch-based system that uses a non-unitary discrimination parameter.

2.5 Conclusion

An analysis based on the ϕ-transformation-framework of Suppes and Zinnes (1963) allows connection of the Rasch model to an absolute difference scale (in log odds form) or a ratio scale (in odds form). When the common discrimination parameter is allowed to vary, it can be classified as an interval scale (in log odds form) or relative ratio scale (in odds form). However, this analysis obscures the constant odds ratio properties of the Rasch model and should only be used if that odds ratio is considered arbitrary, rather than inherent (i.e., in a selected slope paradigm). Meanwhile, the 2PL model can be identified with an interval scale (in log odds form) or relative ratio scale (in odds form).

The theory of Additive Conjoint Measurement provides a possible approach for connecting Item Response Theory Models to scale type theory. However, the latent nature of psychometric attributes and the probabilistic predictions of IRT models prevents the ACM axioms from being directly applicable. Furthermore, the ACM method is designed for purely ordinal data, while IRT data, which contains counts and response frequencies, is richer. Finally, it is important to note that failure to comply with the ACM axioms (as in the 2PL model) does *not* preclude the presence of interval scale properties.

Other derivations of the Rasch model which conclude that it is of an interval scale type have similar flaws. In particular, they fail to consider that appropriate measurement procedures ought to faithfully represent the constant odds ratios suggested by the model, rather than distort them. The choice of 1 as the common discrimination parameter for Rasch model items is not arbitrary, provided the items themselves are not seen as constituting an arbitrary selection from a universe of other, equally valid items, with alternate discriminations.

The conclusions regarding allowable transformations are perhaps the most immediately applicable for practitioners. In the log odds form of the IRT models, the proficiency and difficulty parameters cannot be multiplied by a scalar parameter without a complementary inverse scaling of the discrimination parameter. This fact has been observed before in rescaling of 2PL models (e.g. Davey, Oshima, & Lee, 1996), but has often been neglected by Rasch

practitioners, who rescale model parameters without considering the overall effect on model fit and model predictions.

Chapter 3

Conceptualizing psychometric attributes as quantities or locations

3.1 Introduction

Empirical relationships between physical attributes can often be expressed using simple mathematical operations: a mass that is twice as great, a distance that is half as far. In psychological measurement, we cannot observe such relationships directly. Often, the only data we have are responses to items, which may be scored as correct or incorrect, or perhaps as indicating the presence or absence of an attitude or quality. Alternately, they may be given scores from a set of ordered categories. These responses are then used to assign numbers to respondents according to a selected measurement process. Different processes impose different mathematical structures on the system of assigned numeric values and make different assumptions regarding the nature of the underlying attributes.

The previous two chapters applied scale type theory to psychometric attributes by examining the form of the modeling equation for two commonly used models in Item Response Theory: the Rasch model and the Two-Parameter Logistic (2PL) model. Both models have two (mathematically equivalent) forms of the modeling equation in use, with the logit form being more prevalent. For the Rasch model, the predicted chance of success of an item is given by the following logit equation:

$$\text{logit}(x_i = 1|\theta) = \theta - \beta_i \tag{3.1}$$

where θ represents a respondent's ability and β_i represents an item's difficulty (Rasch, 1960/1980).

Equivalently, it can be modeled using the exponentiated form of this equation, as:

$$\text{Odds}(x_i = 1|t) = \frac{t}{b_i} \tag{3.2}$$

where t represents person ability and b represents the difficulty of item i. All values in Equation 3.2 are exponentiations of the corresponding parameters in Equation 3.1.

Similarly, the related 2PL model has the following logit form:

$$\text{logit}(x_i = 1|\theta) = \alpha_i(\theta - \beta_i) \tag{3.3}$$

where the α_i parameter indicates the discrimination of item i (Birnbaum, 1968). Its exponentiated form is the following:

$$\text{Odds}(x_i = 1|t) = \left(\frac{t}{b_i}\right)^{\alpha_i} \tag{3.4}$$

In the analysis of the previous two chapters, the Rasch model was determined to correspond to either an absolute difference scale (in log odds form) or a ratio scale (in odds form). The Two-Parameter Logistic model was determined to measure attributes along an interval scale (in log odds form) or relative ratio scale (in odds form).

The properties of the two scale types associated with the Rasch model—the absolute difference scale and the ratio scale—are quite distinct. The two scales have different transformations between alternate mappings (Suppes & Zinnes, 1963) and different allowable statistics (Stevens, 1946). The same sorts of differences can be observed in the scales associated with the 2PL model (the interval scale and the relative ratio scale). And yet, both forms of the Rasch model represent in some sense the same model. The forms of the 2PL model are mathematically equivalent, as well. Properties of the attribute being measured using these models do not change depending on the form of the equation.

If there is something lacking in using the Suppes and Zinnes (1963) scale typology (the SZ typology) to analyze models, it is that it obscures the connections between equivalent models, each equally suitable to modeling a type of attribute. It separates the ratio scale and the absolute difference scale, the interval scale and the relative ratio scale, considering them as completely separate scale types, even though no difference can be found in the attributes they are used to measure. The empirical differential isomorphism (EDI) typology used in the first chapter connects the pairs in one way: absolute difference and ratio scales are distinguished by the scalar nature of their empirical differentials, while the empirical differentials of interval and relative ratio scales are classified as magnitudes. But it still falls short of unifying the pairs of scale types completely, or identifying the mathematical properties they share.

Additionally, the SZ typology does not offer any guidelines as to what properties an empirical attribute should have in order to be represented along a certain type of scale or scales. The EDI typology suggests only the nature of the empirical differential, but without a rigorous definition or system for classifying that factor, or any discussion as to what other characteristics are necessary or sufficient for an attribute to be measurable on certain scales. The concept of "magnitudes," in particular, has been left vague thus far, with members of that set identifiable only by the use of units in communicating their size or value.

This paper is in two parts. First, I will discuss properties of types of attributes, using the axiomatic framework of Otto Hölder (1901) . Hölder identifies two types of quantities: magnitudes (Michell & Ernst, 1996), and a second type of quantity which he refers to as points on a line, and which we might call "locations" (Michell & Ernst, 1997). I will present and discuss the axiom sets, and apply them to psychometric attributes, as described by the Rasch and 2PL models.

In the next section, I will look at the properties of scales that can represent these types of quantities, using the homogeneity and uniqueness framework of Narens (1981b). This section will address the question of what features link the absolute difference scale with the ratio scale, the interval scale with the relative ratio scale. It will also use the Rasch and 2PL models as motivating examples.

The two sections have parallel goals. The aim of the first is to apply an alternate framework for classifying psychometric attributes, outside of the traditional "scale types" model. This framework considers not whether attributes are ratio or interval, but whether they have the properties of magnitudes or locations. These properties do not depend on specific parameterizations but are fundamental to the attributes being considered. The goal of the second section is to explore a scale typology which reflects these properties, providing a mathematical basis for scale type assignment which likewise remains consistent across variant parameterizations.

3.2 Types of quantities

3.2.1 Preliminaries

Decades before Stevens (1946), Hölder (1901) introduced two sets of axioms, later translated by Michell and Ernst (1996, 1997), that describe objects or attributes that can be measured along ratio or interval scales. For each axiom, I will examine whether it appears to apply to psychometric attributes, as described by the Rasch and 2PL models. A distinction should be drawn between the following:

1. The respondents;
2. The proficiency of the respondents;
3. A numeric value (the "measurement") assigned to the proficiency of the respondents.

In the Rasch model, θ_A denotes the numeric value assigned to the proficiency a of Respondent A. My goal is to determine whether the proficiencies fulfill the axioms, assuming that there is a map from the proficiencies to the reals such that the Rasch model holds. I will similarly examine the item properties.

Assessing the quantitative nature of attributes involves examining their empirical properties and relations. The Stevens (1946) typology refers to "empirical operations" between

attributes, which are represented by mathematical operations. Under this framework, a nominal scale is one in which there is an empirical operation of "determination of equality." Objects which are empirically determined to be equal or equivalent are then represented with the same number under some measurement map. Objects which can be represented on an ordinal scale must have determination of equality, and "determination of less or greater," a relationship in which the greater object is represented by a greater number. Interval scale attributes have the previous operations, as well as "determination of equal intervals," represented by equality of subtraction. Finally, in order to be represented on a ratio scale, objects must have the operation "determination of equal ratios," represented by equality of quotients. Again, the ratio scale attributes must have all the operations defined by the previous scales.

In psychometrics, the question of what constitutes an "empirical operation" or "empirical relation" is not easily resolved. At the simplest level, respondents who complete an instrument will have some number of items correct. These numbers, like any numbers, can be subject to all sorts of operations: Alice answered twice as many math items correctly as Bob, Bob answered five more correctly than Carol. While performing operations of this type may be appropriate for certain uses, psychometricians would usually prefer to generalize their results beyond performance on a specific instrument. Alice will most likely not answer exactly twice as many items correctly as Bob on the next math test, especially if the items are much easier or harder than those in the precious assessment. Bob answering five more questions correctly than Carol is similarly unlikely to hold up for tests of different difficulty levels, or simply different lengths. However, we may have some expectation that provided both tests are well designed measures of math ability, the order relation is likely (though far from certain) to hold: Alice will probably outperform Bob, who will outperform Carol.

This basic observation—that order relations in terms of raw test scores are fairly stable across instruments targeting the same attribute, while difference and ratio relations are not—invites the conclusion that the psychological attributes, like the sum scores, exist along scales that are, at most, ordinal (Michell, 2009). Much of the work of psychological measurement can be seen as trying to find mathematical relationships, other than order, that are consistent across different items and measures of the same attribute: in essence, placing the respondents on quantitative scales.

The Item Response Theory models suggest that there is a latent value, the probability of responding correctly, that depends on both person ability and item difficulty (as well as item discrimination, for the 2PL model). If this is considered to be an "empirical" value, then it provides a new source of information to use in identifying empirical operations. This approach echoes that of Brogden (1977), who writes:

> For present purposes, which relate to some points of theoretical interest, we will ignore this limitation and assume that precise estimates of the p_{ia} [probabilities] are available, generated through appropriate experimental procedures.

Another way to think of this approach is to consider what the properties of the person

proficiencies and item difficulties would be, if the true success probabilities were to behave according to the Rasch or 2PL model equations. In the place of "empirical operations," I will consider relationships between two respondents that are item-independent, assuming the model's predictions hold. If a relationship between respondents' response probabilities holds regardless of choice of item, it will be acceptable for the axiomatic analysis. Similarly, when considering item properties, I will consider relationships between response probabilities that are respondent-independent. Note that since the alternate forms of each model (logit or odds) result in identical predicted probabilities, these relationships will also be form-independent. This opens the door to the possibility of classifications of attributes that do not depend on parameterization (unlike the EDI or SZ typologies).

A number of these results involve some basic calculations involving the model equations. The details of these calculations are all given in the theorems contained in the Appendix (Section C.3).

3.2.2 The axioms of magnitude

3.2.2.1 Introducing the axioms and connections to the Rasch model

In his first set of axioms, Hölder defines the concept of magnitudes. These axioms, as translated by Michell and Ernst (1996), are as follows.

I. Given any two magnitudes, a and b, one and only one of the following is true: a is identical to b ($a = b$, $b = a$), a is greater than b and b is less than a ($a > b$, $b < a$), or inversely b is greater than a and a is less than b ($b > a$, $a < b$).

The first axiom establishes an order relation, $<$, as a binary operator such that for all a, b either $a < b$, $b < a$, or a and b are identical (denoted by $a = b$). Note that equality itself is not specifically defined, except as occurring when neither $a < b$ nor $b < a$ holds. Hölder mentions in a footnote that he is using equality to denote identical elements, and therefore does not need to specify the usual axioms of equality (If $a = b$ and $b = c$ then $a = c$, etc.).

Additionally, unlike many common definitions of order relations (e.g., Artin, 1991), transitivity ($a < b$ and $b < c$ implies $a < c$) is not given as an inherent characteristic of the order, but will be derived later as a consequence of three of the addition-related axioms. Without transitivity, we are not yet sure that it will be possible to assign numbers to our magnitudes in such away that the greater of two quantities will always be assigned a larger number. (If $a < b < c < a$, no assignment of numerals will preserve this ordering.) Thus, while the determination of equality raises the possibility of applying a nominal scale to our numbers, we cannot yet be sure an ordinal scale would be capable of representing the order relations.

Rasch and 2PL model analysis Both the 2PL model and the Rasch model imply that ordering respondents is possible in an item-independent way (see Theorems 25 and 26). If Alice has a higher probability of success than Bob on one item, then she is predicted to have

a higher probability of success on all items (Figure 3.1). Similarly, if their probabilities of success are the same for a given item, then they are predicted to have equal probabilities of success on all items, and are treated for practical purposes as identical.

On the item side, order is consistent for a Rasch model, but not for a 2PL model (Figure 3.2). This means that Rasch items may be potentially compatible with the axioms of magnitude, but 2PL items are not.

II. For every magnitude there exists one that is less.

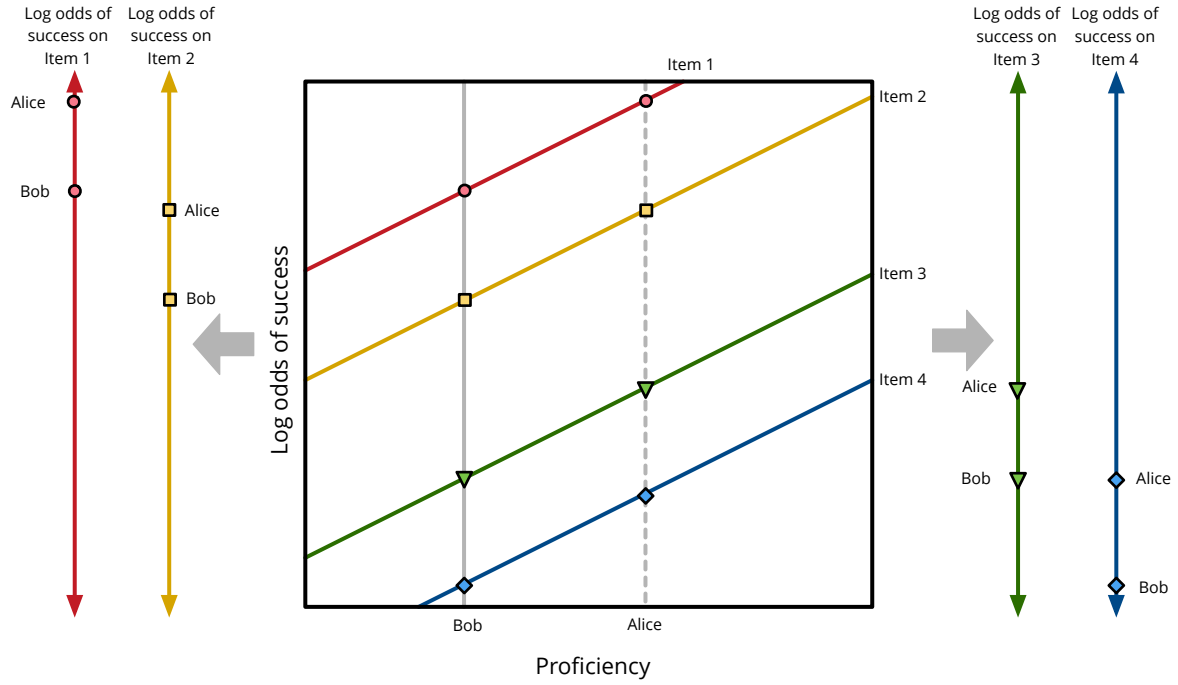
The second axiom states that for every element, there is one that is smaller. If we had transitivity, we could use this axiom to establish that our set of magnitudes must not be finite. Without transitivity, cycles such as $a < b < c < a$ will require additional axioms to eliminate.

Rasch and 2PL model analysis For psychometric data, this axiom is not possible to verify empirically, even in the imagined case where all latent probabilities are known. In a finite population, it is arguably strictly false, since a well-ordered finite group must have minimal elements. We could ask instead, given any respondent, is there a lesser amount of ability that is possible to have? Or, on the item side, given any item, is it theoretically possible to design an easier item? From a theoretical perspective, these questions could arguably be answered in the affirmative. While any particular instrument has a lowest possible score (zero on all items), the attribute being measured need not have a theoretical minimum proficiency, nor easiest item. In theory, an item could be added to the assessment which is easier than all the current items, lowering the minimum possible proficiency which could be measured by the assessment.

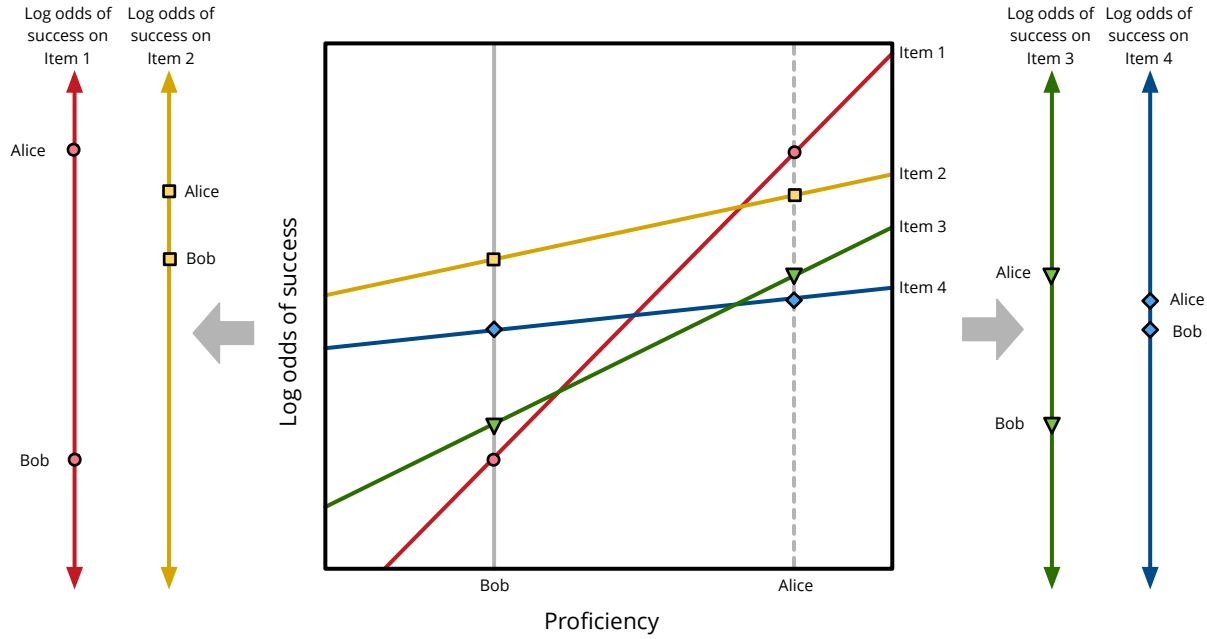
III. For every ordered pair of (not necessarily distinct) magnitudes, a and b, their sum, $a + b$, is well-defined.

The third axiom establishes an addition function, $+$, under which the set of quantities must be closed. As in Axiom I, the '$=$' of $a + b = c$ is undefined. Hölder's notes confirm that he considers "equal added to equal results in equals" (presumably, that $a = n$ and $b = m$ implies $a + b = n + m$) to be an unnecessary axiom that is an obvious consequence of considering equal magnitudes to be identical. This axiom establishes only that any two elements have a sum, without any further restrictions on what that might entail.

Addition is often connected to the concept of *concatenation*. The usual system of distance measurement ensures that the sum of the numbers assigned to the lengths of two rods is equal to the number assigned to the length of the rods concatenated together. Mass and weight behave similarly. Other physical quantities, such as temperature or density, are more difficult to add directly.

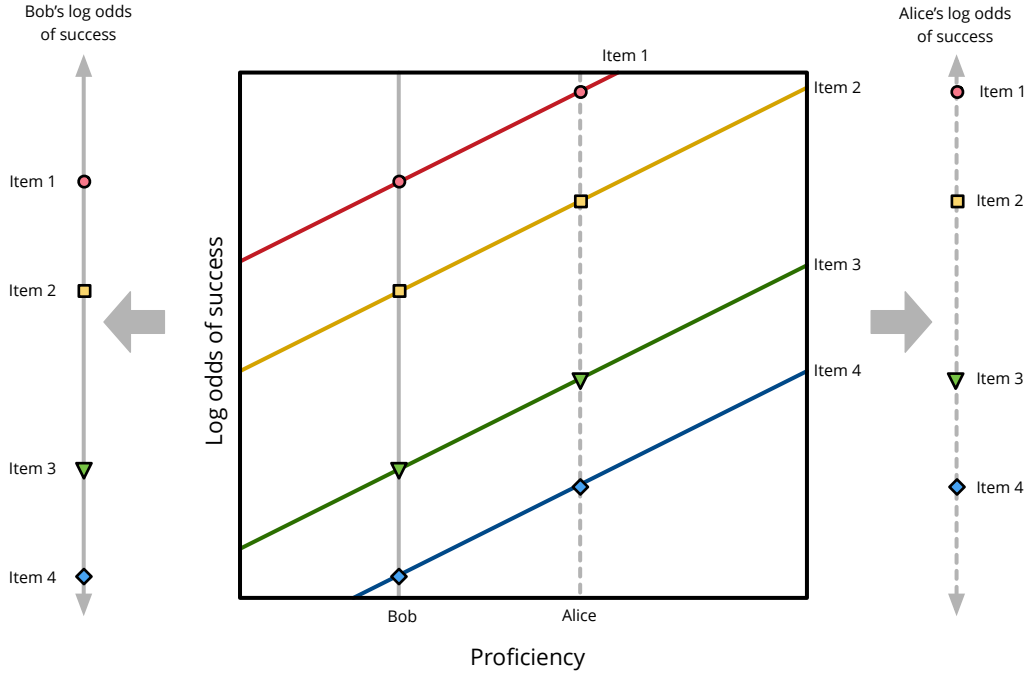


(a) In a Rasch model, Alice's chances of success on each item are higher than Bob's, so person order is consistent across items.

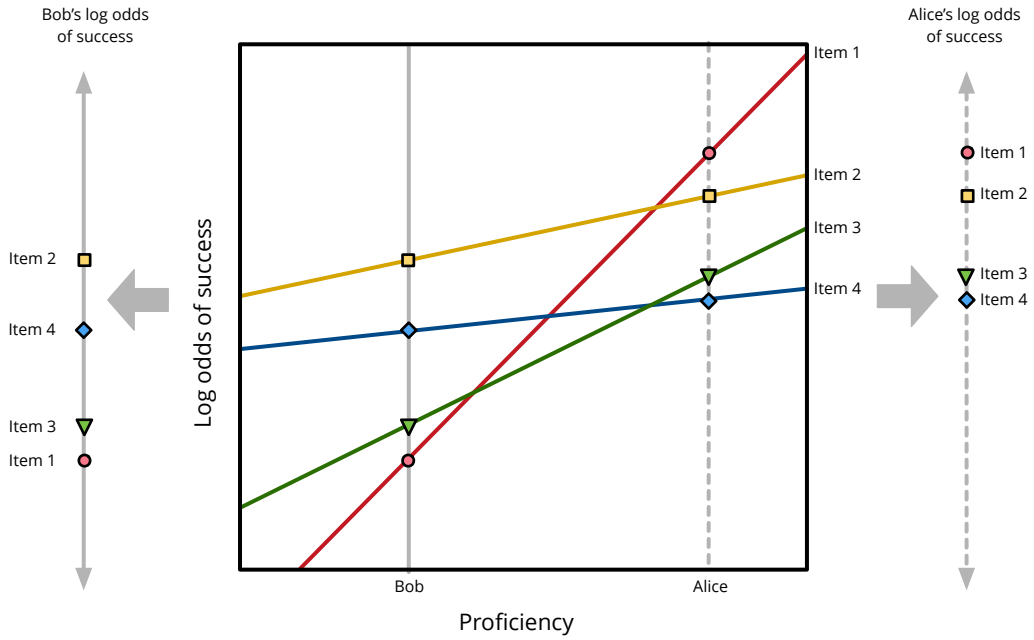


(b) In a 2PL model, Alice's chances of success on each item are higher than Bob's, so person order is consistent across items.

Figure 3.1: Person order in Rasch and 2PL models.



(a) Items in a Rasch model have the same order for all respondents.



(b) Items in a 2PL model have different orders for different respondents.

Figure 3.2: Item order in Rasch and 2PL models.

Rasch and 2PL model analysis In psychometrics, it is not generally considered possible to empirically concatenate or combine two respondents. One idea might be to add together values associated with respondents' chance of success on items. For example, if Alice's predicted probability of success on an item is the sum of Bob's probability and Carol's probability, we might imagine claiming Alice's ability to be the sum of Bob's and Carol's abilities. The downside of this proposal is that for neither the Rasch model nor the 2PL model is this relationship item-independent; if it holds for one item, it will fail for an item of a different difficulty level.

However, there is a related value for which the summation relationship is independent of the item difficulty: odds. The odds form of the Rasch model was given in Equation 3.2. If Alice's *odds* of success on an item are the sum of Bob's and Carol's odds on that item, this will be true for any item included in the model. If for some item i we have a summation relationship, it will hold for any other item j (see Theorem 27). We now have an item-independent way to consider Alice's ability to be the sum of Bob's and Carol's abilities. The same relationship holds on the item side: If Alice's odds of success on item i is the sum of her odds of success on items j and k, then the same relationship will hold for Bob, or any other respondent (Theorem 28).

These relationships do not hold for the 2PL model, and in fact there does not appear to be a simple method for defining summation in this model in an item-independent way (on the person side) or a person-independent way (on the item side). This result suggests that abilities and items behaving according to the predictions of the 2PL model do not conform to Hölder's axioms. However, it is still possible for abilities and items performing as predicted by the Rasch model to be compatible (pending analysis of the remaining axioms), since they do appear to have an empirical summation relationship. Note, though, that just as a finite dataset may be seen as technically violating Axiom II by having a minimal element, a real dataset may not directly satisfy Axiom III by containing a respondent whose predicted odds of success are equal to exactly the summed odds of two given respondents. In fact, even a large dataset may contain exactly zero sets of three respondents in which the odds of success of two of the respondents exactly sum to the odds of the third. This, then, is again a question of an axiom holding true in a theoretical sense rather than a practical case. That is, while there *could* be such a person, one is not guaranteed.

IV. $a + b$ is greater than a and greater than b.

The fourth axiom, which states that the sum of two numbers is always greater than either, is the first that is not satisfied by the set of real numbers together with the normal addition operation, since adding a negative results in a lesser number, not a greater. To satisfy this axiom, we could restrict the set to positive numbers, or redefine '+' to denote "The sum of the squares." This axiom, together with Axiom III, implies that there is no greatest element of the set, since any addition operation will yield one that is greater.

Rasch and 2PL model analysis Since odds are always positive, this axiom is easily satisfied by using the odds summation definition for addition as discussed in the previous axiom.

V. If $a < b$, then there exist x and y such that $a + x = b$ and $y + a = b$.

The fifth axiom implies what we might refer to as "subtraction." Given elements b and a such that $b > a$, we can find x such that $a + x = b$ and y such that $y + a = b$. (Note that since we have not yet established commutativity, we cannot assume that these are equal.) This axiom is a strong step toward narrowing the set of potential operations for '+'. (This eliminates the reals + "sum of squares" possibility discussed in Axiom IV. For example, if b and a are both negative with $a < b$, then there is no real x such that the sum of the squares of a and x will yield b.)

Rasch and 2PL model analysis Like Axioms II and III, this existence axiom is unlikely to hold for real data, but is certainly possible in principle for data obeying a Rasch model. This axiom is very similar to the solvability axiom from ACM. In this case, it states that if Bob's chances of success are higher than Alice's, then there exists a level of ability such that for any respondent Xavier at that ability level, Bob's odds of success are equal to Alice's odds added to Xavier's odds. We saw from the discussion in Axiom I that if Bob's odds of success are higher than Alice's on one item then they are higher on all items. From the discussion in Axiom III, we know that if Bob's odds of success on one item are equal to the sum of Alice's and Xavier's odds of success, then this relationship holds for all items. A parallel analysis is possible for items in a Rasch model. In principle, there is no reason to suppose that a particular level of ability is impossible for a respondent to hold, or that it is impossible to construct an item with a particular level of difficulty, although it is likewise not guaranteed.

VI. It is always true that $(a + b) + c = a + (b + c)$.

While still not establishing commutativity as an axiom, Hölder's sixth axiom defines addition as associative. In our running psychometric example, adding odds clearly obeys this axiom.

VII. Whenever all magnitudes are divided into two classes such that each magnitude belongs to one and only one class, neither class is empty, and any magnitude in the first class is less than each magnitude in the second class, then there exists a magnitude a such that every magnitude b such that $b < a$ is in the first class and every magnitude c such that $c > a$ belongs to the second class. (Depending on the particular case, a may belong to either class.)

The last axiom establishes that our magnitudes form a linear continuum. This axiom is trivially true in a finite set of psychometrics data, and true by properties of the real numbers in the set of all possible odds values.

Hölder uses these axioms to establish transitivity of the order relation, density of the order, uniqueness of subtraction, commutativity of addition, and the Archimedean axiom. Transitivity is the last piece required to make Hölder's order relation potentially isomorphic to the standard order relation on the reals, introducing the possibility of applying an ordinal scale to these quantities. Since these properties are shown to hold as a result of the axiom set, there is no need to verify their correspondence with psychometric attributes.

Based on the preceding analysis, attributes that behave as predicted by the Rasch model are an example of magnitudes, as defined by Hölder. This analysis did not require assigning numbers to the proficiencies, and did not depend on any specific parameterization. Unlike the scale types of interval, ratio, etc., Hölder's classification of "magnitude" applies to properties of the psychometric attribute, regardless of choices made in parameter assignment.

3.2.2.2 The scale type of magnitudes

The traditional scale types are still useful in several ways. They apply more concretely to the actual measurement map used, constraining the allowable transformations and statistics that can be used with a set of measurement data. One advantage of using Hölder's "magnitude" classification is that if the scale types applicable to magnitudes are known, then any property shown to be compatible with the axioms of magnitude can automatically be known to be measurable along those types of scales.

Under the EDI typology, in order to successfully apply ratio or absolute difference scale to magnitudes, a function will be needed that defines an empirical difference relation between magnitudes with a real-valued empirical differential which is isomorphic to the mathematical notion of ratio or subtractive difference. For this, I will use Hölder's *measure-numbers*.

Axiom III defined the addition of two magnitudes. From addition, Hölder inductively defines scalar integer multiplication, starting with

$$2a = a + a \tag{3.5}$$

and continuing to

$$na = a + (n-1)a. \tag{3.6}$$

Multiplication can be used to define division. Ideally, it would be nice to define the ratio $[a : b]$ as the real number κ such that

$$a = \kappa \cdot b \tag{3.7}$$

However, multiplication has only been defined for integer multiplicands, and there is no reason that there should exist an integer κ to make this true. A rational measure-number $[a : b]$ could be potentially be defined as $\frac{\mu}{\nu}$ (with integer μ, ν) such that

$$\nu \cdot a = \mu \cdot b \tag{3.8}$$

but there is no reason that there should exist a rational number with this property either. If, for example, a and b represent the lengths of the side and diagonal of a square, or the diameter and circumference of a circle, they will not be relatable rationally.

However, if all the rational numbers $\frac{\mu}{\nu}$ are divided into two sets, depending on whether $\nu a < \mu b$ or $\nu a > \mu b$, the upper limit of the smaller set and the lower limit of the greater set will be equal. This limit can then be defined as the measure-number $[a : b]$. In other words, the measure-number represented by $[a : b]$ is defined as the real number such that

- For all rational numbers $\frac{\mu}{\nu}$ less than $[a : b]$, $\nu a > \mu b$; and
- For all rational $\frac{\mu}{\nu}$ greater than $[a : b]$, $\nu a < \mu b$.

Because $[a : b]$ partitions all rational numbers into two classes depending on whether they are less than or greater than the measure-number, it is also referred to as a *cut*. In the EDI typology, this constitutes an empirical difference relation with a scalar valued empirical differential.

Just as in the EDI typology I used odds ratio as the empirical differential in a a Rasch model, odds ratio can be used here to define the measure-number between two respondents. This value is item-independent in a Rasch model (see Theorem 29), but not in a 2PL model.

Hölder derives several properties for measure-numbers. The ones that are most relevant for the following sections are given in the Appendix (Section C.1).

Measure-numbers can be used to construct a ratio scale on Hölder's quantities. For the rule in assigning numerals to quantities, use the following procedure:

Procedure 1.

1. Select some quantity b as the *unit*.
2. For each quantity a, assign a the measure-number $[a : b]$.

For psychometric quantities as modeled by the odds form of the Rasch model as given in Equation 3.2, this procedure becomes:

Procedure 2.

1. Select some respondent B. Assign a proficiency value $t_B = 1$.
2. For each respondent A, assign to proficiency parameter t_A the ratio of A's odds of success on an item and B's odds of success on an item. By Theorem 29, this value is independent of the item selected.

If item difficulties are then assigned in the following way:

$$b_i = \frac{1}{\text{Odds}(x_i = 1|t_B)} \tag{3.9}$$

then the full Rasch model holds (Theorem 30). Proficiencies under this map meet the definition for a ratio scale according to the Suppes and Zinnes (1963) framework, based on ϕ transformations (Theorem 31).

In the previous chapter, it was demonstrated that the odds form of the Rasch model given in Equation 3.2 establishes a ratio scale. It was also noted that the log odds form of the model establishes an absolute difference scale. Analogously, an alternate mapping procedure exists for Hölder's magnitudes for which the maps form an absolute difference scale. This procedure is as follows:

Procedure 3.

1. Select some quantity b to have the value 0.
2. For each quantity a, assign a the log of the measure-number $[a : b]$.

See Theorem 32 for a demonstration that the set of ϕ transformations between mappings following Procedure 3 are exactly the set of functions involving addition of a constant, making this an absolute difference scale. Note also that if Procedures 1 and 3 use the same quantity b for Step 1, then the values assigned under Procedure 3 will be the log of the values assigned under Procedure 1, for all elements, making the two maps equivalent under a log transformation.

The psychometric equivalent of Procedure 3 is as follows:

Procedure 4.

1. Select some respondent B. Assign a proficiency value $\theta_B = 0$.
2. For each respondent A, assign to proficiency parameter θ_A the log of the ratio of A's odds of success on an item and B's odds of success on an item. By Theorem 29, this value is independent of the item selected.

This procedure assigns respondents the log of the proficiency values assigned through Procedure 2, and establishes an absolute difference scale.

Together, these results show that in principle, if odds of success are treated as observable, and they conform to the assumptions of the Rasch model, then the underlying attribute meets Hölder's definition of a magnitude. It has also been shown that these magnitudes can be measured along either a ratio or an absolute difference scale, and that those pairs of

scales are equivalent under a log transformation. However, we were not able to apply the same approach to the 2PL model. In the next subsection, I will apply the same methods in considering the 2PL model in conjunction with Hölder's axioms of points on a line.

3.2.3 The axioms of points on a line

.

3.2.3.1 Introducing the axioms and connections to the 2PL model.

Hölder's second set of axioms define the concept of *points* and the related concept of *intervals* (Michell & Ernst, 1997). For notation, Hölder uses capital letters A, B to represent points, and pairs of points AB to represent the interval from point A to point B.

Hölder's axioms can be classified into a few categories:

1. Axioms concerning order of points (alternately expressed by Hölder as the "direction" of a given interval);
2. Axioms of equivalence of intervals;[1]
3. Axioms concerning additional properties of intervals;
4. Axioms of existence, continuity, and density of points.

Throughout the paper, Hölder presents a number of axioms, some of which imply each other. The set given here, labeled with the Greek letters given in the original, is one Hölder specifically identifies as non-redundant. In this set, the axioms concerning point order are (Michell & Ernst, 1997, p. 346):

(β) Intervals within a straight line are of two kinds, such that any interval is of one and only one kind. Intervals of the same kind are called "of the same direction," and intervals of different kinds are called "of opposite direction." The intervals AB and BA are always of opposite direction. Let the intervals of one kind be called "intervals of the first direction" and the fact that AB is an interval of the first direction be expressed as $A \subset B$ or $B \supset A$.

(γ) From $A \subset B$ and $B \subset C$ it always follows that $A \subset C$.

These axioms establish a general order relation. Unlike for the axioms of magnitude, transitivity is established right away.

[1]Note that Hölder does not explicitly include equivalence axioms of points. Throughout the axioms, Hölder will treat points as objects that are either distinct or identical, with the properties of that identity assumed.

Rasch and 2PL model analysis As discussed in the previous section, it is possible under either a Rasch model or a 2PL model to order respondents by predicted chance of success, independent of which items they are given (Figure 3.1; Theorems 25 & 26). Then it can be said that "Alice-Bob" is an interval of the first kind if Alice's predicted probability of success on any item is less than Bob's. However, while items in a Rasch model can be ordered independently of persons, items in a 2PL model cannot (Figure 3.2). This will again proclude items in a 2PL model from conforming to this set of axioms. 2PL proficiencies, however, are not eliminated from consideration by this axiom.

The axioms concerning equivalence of intervals are (Michell & Ernst, 1997, p. 349):

(μ) Any two intervals, be they of the same or opposite direction, can be compared such that they are found either equal or unequal in a specific manner.

(ν) Two intervals each equal to a third are equal to one another.

Axiom (μ), which states that any two elements are either equal or are not equal, is usually treated in logical frameworks as a tautology rather than as an axiom. Here, this axiom functions as the first indication that the intervals AB, BA, etc., have more properties than simply an ordinal direction, but also have values that may be equal or unequal. These values can also be thought of as the *lengths* of the intervals, where intervals of the same length constitute an equivalence class.

Another key clause here is that the intervals *can be compared*. This recalls Stevens' description of these relations as empirical, rather than abstract.

Rasch and 2PL model analysis In psychometrics, "distance" between respondents is not an immediately obvious trait. In the Rasch model, the difference in log odds between two respondents is independent of the item (see Theorem 33). This means that if we think of the length of the interval between two respondents as the size of their odds ratio, this constant value is item independent. Similarly, the difference in log odds between two items is independent of the respondent (Theorem 34).

For the 2PL model, the difference in log odds between two respondents on an item depends on the item's discrimination (Theorem 35). However, there is nothing in Axioms (μ) or (ν) to require the length to be a numeric constant. It is only necessary for lengths to be comparable, and found either equal or unequal. This property is independent of the item chosen. If two pairs of respondents have the same difference in log odds on one item, they will on all items (Theorem 37).

Hölder introduces another equivalence axiom for intervals:

(*) [2]

> AB is always equal to BA, even if not identical to it (Michell & Ernst, 1997, p. 355).

[2]Hölder does not assign a letter to this axiom, but refers to it simply as $AB = BA$.

Axiom (*) establishes that intervals AB and BA are in the same equivalence class and exhibit the same length. In psychometrics terms, he is establishing "twice the odds" and "half the odds" as equal, in opposite directions.

In a preliminary sense, the above axioms define the required relations for an interval scale per Stevens (1946): determination of equality (implied by identity), determination of lesser or greater (Axioms (β) and (γ)), and determination of equality of intervals (Axioms (μ) and (ν)). However, without further exploration of the properties of the intervals and how they connect to the ordinality axioms, it is not guaranteed that there will be an appropriate isomorphic measurement map to the reals available. Suppose, for instance, that all intervals were defined to be equal to each other. In a trivial sense, this would satisfy all the above axioms. However, any map to the reals attempting to keep equality of intervals isomorphic to equality of differences (while maintaining isomorphism of equality and order of points) would be doomed to failure. Further axioms will constrain the properties of intervals appropriately.

The axioms describing further properties of intervals include:

(θ) If $M \subset N$ and A is an arbitrary point then there exists exactly one point B such that $A \subset B$ and $AB = MN$ and exactly one point C such that $C \subset A$ and $CA = MN$.

(o) If $A \subset B \subset C$ and $A' \supset B' \supset C'$, then if $AB = A'B'$ and $BC = B'C'$ then $AC = A'C'$.[3]

Intuitively, Axiom (θ) can be thought of as describing "moving" an interval to a different end point. Axiom (o) can be thought of as concatenating two intervals, and states that this stays consistent even if

- The intervals are replaced by other members of their length equivalence class;
- The replacement intervals are of the opposite direction and order.

Rasch and 2PL model analysis As in the existence axioms from the axioms of quantity, Axiom (θ) is an existence axiom which may be true in principle for a Rasch or 2PL model (there is exactly one ability level which will fit the desired property) but which may not actually exist in a given data set. Keeping length as "difference in log odds," Axiom (o) holds for the Rasch and 2PL, and is item independent.

The last group of axioms concern properties of existence, density, and continuity:

(α) On a straight line there are at least two distinct points.

[3]Michell's 1997 translation gives the condition as $A \subset B \subset C$ and $A' \subset B' \subset C'$. However, this would make (o) identical to an earlier axiom (η). Furthermore, the notes referring to (o) are only comprehensible if the order of the two sets is reversed. Referring to the original German edition of the axioms (Hölder, 1901) shows that the formulation given here is correct, and the Michell edition is in error.

(δ) If $A \subset C$ then there exists at least one point B such that $A \subset B$ and $B \subset C$. It is then said that the point B lies "between" A and C or that it lies between C and A.

(κ) If all the points on a straight line are divided into two classes such that neither class is empty, each point belongs to exactly one class, and for every point X in the first class and every point Y of the second class $X \subset Y$, then there exists a point Z such that each point $A \subset Z$ belongs to the first class and each point $B \supset Z$ belongs to the second class.

Rasch and 2PL model analysis The first of these axioms eliminates trivial sets, and should be true for any non-trivial IRT model. The second axiom establishes a dense order, while the last axiom, concerning Dedekind cuts (Dedekind, 1901), is an axiom of continuity. These last two axioms are crucial to the understanding of an interval as continuous, but are less relevant to the discussion of the distinction between different scale types. For a finite set, Axiom (δ) will be false, while Axiom (κ) will be true. However, both can be taken to hold in principle for a Rasch or 2PL model, if all possible person proficiencies are considered.

Thus, proficiency attributes as modeled by both the Rasch model and the 2PL model conform to the axioms of points on a line, as do item difficulties as modeled by the Rasch model. Proficiencies under the 2PL model, which did not conform to the axioms of magnitude in Section 3.2.2.1, may be a natural fit for the location axioms. However, it may seem curious that the Rasch model fits both sets of axioms. Are proficiencies and items under this model better thought of as magnitudes, or as locations?

In fact, as demonstrated in Theorem 38, all attributes that comply with the axioms of magnitude automatically fulfill the axioms of points on a line. This echoes the Stevens hierarchy of scale types, in which a higher scale type (such as ratio) subsumed all the properties of the lower scale types (such as interval). Similarly, in the Suppes & Zinnes framework, ratio scales are those for which the ϕ transformations are scalar multiplications, while for interval scales, the set of ϕ transformations are linear transformations. All the properties of interval scales hold true for elements measured on ratio scales—but not necessarily vice versa. Similarly, magnitudes have all the properties of locations, but attributes that function as locations are not necessarily magnitudes.

While the axioms of magnitude can be applied to locations, there is a more natural application of them within the framework of the points on a line. As shown by Hölder (Michell & Ernst, 1997), the distance between locations can be itself considered a magnitude (see discussion in Theorem 39). If distances of the same direction a and b can be defined as magnitudes, then the measure-numbers $[a : b]$ can also be defined. For simplicity, the notation $[AB : CD]$ will be used to denote the ratio between the distances exhibited by AB and CD respectively (defined only if they are of the same direction).

3.2.3.2 The scale type of locations

Numeral assignment Just as it was shown in Section 3.2.2.2 that Hölder's magnitudes can be represented by a ratio scale, so too can an interval scale be applied to the locations defined by the point and interval axioms. Hölder gives a procedure for assigning each point a number as follows:

Procedure 5.

1. Select an arbitrary point N to be assigned the numeral 0.
2. Select an arbitrary point E to be assigned the numeral 1.
3. For any other point A, if NE and NA are of the same direction, assign to A the measure-number $[NA : NE]$. If NE and NA are of opposite directions, then NE and AN will be of the same direction. Assign to A the negative of $[AN : NE]$.

Following this procedure, the order of the numbers assigned to any two magnitudes A and B will depend on whether $N \subset E$ or $E \subset N$: If AB and NE are the same direction, the number assigned to A will be less than that assigned to B; otherwise B will be assigned a greater number. Since the Suppes and Zinnes (1963) specifications for any scale of type ordinal or higher require that the order relation remain constant under the class of possible numeral assignment rules, this procedure must be modified slightly in order to conform to their definition. The following procedure accommodates these requirements:

Procedure 6.

1. Select an arbitrary point N to be assigned the numeral 0.
2. Select a point E such that $N \subset E$ to be assigned the numeral 1.
3. For any other point A, if $N \subset A$, assign to A the measure-number $[NA : NE]$. If $A \subset N$, assign to A the negative of $[AN : NE]$.

For a psychometric quantity, this procedure can be expressed as follows for the logit version of the 2PL model (Equation 3.3):

Procedure 7.

1. Select two arbitrary respondents E and N with different predicted odds of success on any item. By Theorem 26, these respondents can be ordered such that if E has a higher predicted odds of success than N on one item, then their odds of success are higher on any item. Assume without loss of generality that E has higher predicted odds of success than N.

2. Assign the value $\theta_N = 0$ to N and the value $\theta_E = 1$ to E.

3. For any other respondent A, the following will be a constant value across all items i (Theorem 36):
$$\frac{\text{logit}(x_i = 1|\theta_A) - \text{logit}(x_i = 1|\theta_N)}{\text{logit}(x_i = 1|\theta_E) - \text{logit}(x_i = 1|\theta_N)} \tag{3.10}$$
Assign this value to θ_A.

To complete the model, assign to Item i the following discrimination parameter:

$$\alpha_i = \text{logit}(x_i = 1|\theta_E) - \text{logit}(x_i = 1|\theta_N) \tag{3.11}$$

and the following difficulty parameter:

$$\beta_i = -\frac{\text{logit}(x_i = 1|\theta_N)}{\text{logit}(x_i = 1|\theta_E) - \text{logit}(x_i = 1|\theta_N)}. \tag{3.12}$$

Parameters assigned in this way are consistent with the 2PL model (Theorem 40).

The points on a line, as defined by these axioms, meet the requirements of the Suppes and Zinnes (1963) definition of an interval scale; namely, that all the possible mappings under these rules are distinguished by a positive linear transformation (Theorem 41). Since the proficiencies as modeled by a Two-Parameter Logistic model conform to these axioms, they can also be represented on an interval scale.

In the previous chapter, it was shown that the 2PL proficiencies can also be represented on a relative ratio scale using the odds form of the model. This practice is not common; the log odds form of the model is more commonly used. However, for the sake of completeness, and to illustrate some commonalities between different assignment procedures for the same attributes, the following algorithm represents a mapping of locations onto a relative ratio scale:

Procedure 8.

1. Select an arbitrary point N to be assigned the number 1.

2. Select an arbitrary point $E \supset N$ to be assigned the number e.

3. For any other point A, if $N \subset A$, assign the number $\exp([NA : NE])$. If $N \supset A$, assign the number $\exp(-[AN : NE])$.

For mapping functions defined by this procedure, the set of ϕ transformations is the set of functions of the form $\phi(x) = \gamma \cdot x^{\alpha}$, meaning that a relative ratio scale can also be defined on locations (Theorem 42).

3.2.4 Summary

Attributes under the Rasch model conform to the axioms of quantity and are therefore ratio-scaleable (or represented by absolute differences), while attributes under the 2PL model conform to the axioms of points on a line and can therefore be represented with interval or relative ratio scales. These conclusions depend on accepting the assumptions, for example, that there is no minimum element in a Rasch model, or that there can always be found a respondent whose proficiency as modeled by the 2PL model is the desired distance from another respondent. These assumptions are not contradictory to properties of these models. They resemble the some of the axioms of Additive Conjoint Measurement (for example, solvability) which were discussed in the previous chapter and which are commonly said to hold for the Rasch model.

This classification into magnitudes and points on a line provides a way to think about types of attributes in a way that is independent of the specific numerical mapping applied, and demonstrates that ratio scales and absolute difference scales can be applied to a certain type of quantity, while interval scales and relative ratio scales can be applied to another. The next section will discuss connections between these pairs of scales that can be defined on the same attributes, and provide an overarching scale type framework for defining these types of sister scales.

3.3 Mapping choices

3.3.1 Generalized maps

The previous section established that magnitudes could equivalently be represented using a ratio scale or an absolute difference scale, while points on a line could be represented using an interval scale or a relative ratio scale. These are in every sense equivalent maps (related through exponentiation), yet the typology we have been using to categorize them keeps them distinct. This typology is useful for discussing properties of the assigned numeric values, but less useful for classifying the underlying quantities or understanding which types of maps may be useful. In this section, I will discuss an alternate typology system that groups together these types of equivalent maps, and discusses the mathematical properties they have in common.

Procedure 6 in Section 3.2.3.2 assigned numeric values to the points on a line. This involved selecting one point to be assigned the value 0, and another to be assigned the value 1. There is nothing special about using the numbers 0 and 1 for this purpose, except that they are easy to work with. Procedure 12 defines a general version of this process which

allows arbitrary numeric values to be assigned for the first two points (Appendix, Section C.2.1). Like Procedure 6, it defines an interval scale (Theorem 43).

The algorithm in Procedure 8, defining a relative ratio scale, can be generalized as well to allow the choice of any two arbitrary positive real numbers to be mapped to any pair of elements (Procedure 13, Appendix). It defines a relative ratio scale (Theorem 44).

This gives us two possible maps on points on a line for which the empirical relation of equality of intervals is isomorphic to a given difference relation. These mapping procedures have some properties in common, including the following:

- Up to an order relation, the processes begin by selecting any two points M and N and mapping them to any two numbers (in the appropriate range).
- After fixing the images of these points, the definition of the rest of the map follows with no further choices.

These definitions echo the establishment of common interval scales, such as the Celsius scale for temperature: Assign 0 to the freezing point of water, and 100 to the boiling point, and the rest of the scale is defined.

In the ratio scale case, Procedure 1 defined a ratio scale on magnitudes by first selecting an element to serve as the unit (assigned a value of 1). A more generalized version of this procedure is given in Procedure 14 (Appendix, Section C.2.1), in which any one element is assigned any (positive real) value. It establishes a ratio scale (Theorem 45). Similarly, Procedure 3, which chose a quantity to be assigned the value zero, can be generalized to assign any value to its initial element (Procedure 15, Appendix). This procedure establishes an absolute difference scale (Theorem 46). These procedures are also structurally similar: They begin by assigning any value to any one point. The rest of the scale is then automatically defined.

Narens (1981b) defined a *relational structure* $\mathscr{X} = \langle X, R_0, R_1, \ldots \rangle$ as a set X along with its relations $(R_0, R_1, \ldots)$. Let $\mathscr{N} = \langle N, S_0, S_1, \ldots \rangle$ be a relational structure with $N \subseteq \mathbb{R}$. Then a measurement map can be thought of as an isomorphism $f : \mathscr{X} \to \mathscr{N}$ that preserves the relational structure. For given $\mathscr{X}, \mathscr{N}$, there may be many such possible maps. This introduces the idea of choice: For a given $x \in X$, we may be able to choose a map f such that $f(x) = p$, or map g with $g(x) = q$. The question then arises as to how much choice we have in making these assignments. For this section, I will follow Narens (1981b) in only considering structures in which the first relation R_0 is a total order relation $\succeq$. Because the properties presented here are closely linked, I will present both before discussing the connection to psychometric variables.

Homogeneity The concept of n-point homogeneity (Narens, 1981b) can be defined as follows:

> Let $\mathscr{X} = \langle X, \succeq, R_1, R_2, \ldots \rangle$ and $\mathscr{N} = \langle N \subseteq \mathbb{R}, S_0, S_1, \ldots \rangle$ be relational structures, and let F be the set of isomorphic measurement maps from $\mathscr{X}$ to $\mathscr{N}$.

> Then $\mathscr{X}$ satisfies n-point homogeneity if and only if for any ordered sets $x_1 \succ x_2 \succ \ldots \succ x_n$ and $p_1 \succ p_2 \succ \ldots \succ p_n$ with $x_i \in X$ and $p_i \in N$, there exists a map $f \in F$ such that $f(x_i) = p_i$ for all $1 \leq i \leq n$.[4]

In other words, for relational structures with n-homogeneity, any set of n attributes can be mapped to any set of n numbers, up to an order restriction. It follows immediately from the definition that any relational structure for which the set of isomorphisms satisfies n-point homogeneity will also satisfy k-point homogeneity for all whole numbers k smaller than n. The maximum n such that $\mathscr{X}$ satisfies n point homogeneity can be referred to as the degree of homogeneity of $\mathscr{X}$. Procedures 14 and 15, which began by assigning arbitrary numbers to any one point, illustrate the 1-point homogeneity of magnitudes. Procedures 12 and 13, which began by choosing any two elements and assigning any pair of numbers, thereby demonstrate that the underlying structures (the points on a line) have 2-point homogeneity. Structures with finite homogeneity values greater than 2 are rare.

If $\mathscr{X}$ satisfies n-point homogeneity for all possible integer values of n, we say that $\mathscr{X}$ is ∞-point homogeneous (Narens, 1981b). These structures are represented by ordinal scales.

Uniqueness For the second property introduced by Narens (1981b), n-point uniqueness, we can use the following definition:

> Let $\mathscr{X} = \langle X, \succeq, R_1, R_2, \ldots \rangle$ and $\mathscr{N} = \langle N \subseteq \mathbb{R}, S_0, S_1, \ldots \rangle$ be relational structures on X and $\mathbb{R}$ respectively, and let F be the set of isomorphic measurement maps from $\mathscr{X}$ to $\mathscr{N}$. Then $\mathscr{X}$ is said to satisfy n-point uniqueness if, for any two maps $f, g \in F$, if there is a set of n elements $x_1, x_2, \ldots, x_n \in X$ such that $f(x_i) = g(x_i)$ for all $1 \leq i \leq n$, then $f(x) = g(x)$ for all $x \in X$.

Whereas n-point homogeneity concerned whether a given set of elements could be mapped to a given set of numbers, n-point uniqueness focuses on whether knowing the images of a given set of points determines the entire map. A relational structure that satisfies n-point uniqueness will also satisfy k-point uniqueness for all whole numbers k greater than n, since if n fixed points are always enough to determine a map, $n + 1$ points will be more than sufficient. The degree of uniqueness of the structure can be defined as the minimum n such that $\mathscr{X}$ is n-unique. Procedures 12 and 13, under which the map was completely defined after the choice of two points, demonstrated the 2-point uniqueness of the points on a line, while Procedures 14 and 15, under which the map was determined after the choice of a single point, demonstrate the 1-point uniqueness of Hölder's magnitudes.

[4] This is not quite the definition given in Narens (1981b). He gives n-point homogeneity, and the related property n-point uniqueness, as attributes of the set H of automorphisms from $\mathscr{X}$ to itself. However, as he notes, the set of such automorphisms is directly related to the set of measurement maps, and the two sets have equivalent properties. In this paper, I choose to focus on the interpretation that concerns the measurement maps as most relevant for the discussion.

A note: the uniqueness property described here, in which a line is "defined" by the choice of two points and their numeric values, should not be confused with the geometric property in which a line in space can be "defined" by identifying two of its points. In the first case, the task at hand is to define the *scale* of the line by assigning a single number to each point along it. In the second case, the task at hand is to determine *which* points in space are or are not on the line. "Identifying" two points in this geometric case consists only of noting two points which are on the line. In the uniqueness property case, by contrast, there is no question of which points are on the line, only of what the scale of the line should be.

This contrast can also be seen by examining the case of the plane. In geometry, three points "define" a plane, which is to say that identifying any three (non-collinear) points as belonging to a plane is sufficient to classify all other points as on or not on the plane. However, the uniqueness and homogeneity of the points on a plane are undefinable in this system, since these properties are only defined for ordered structures in which each element is assigned a single point. Points in a plane, being multidimensional, do not have a strict order and individually require multiple coordinates to define. A similar issue exists with regards to the scale type of lines in a plane (discussed more later in with reference to items in a 2PL model).

When the map is not uniquely determined regardless of the size n of the set of known points, we say that the structure has ∞-point uniqueness. Note that this does not necessarily imply ∞-point homogeneity, as not all assignments may be possible. Alternately, a structure with 0-point uniqueness is one for which there is only a single absolute map, and thus no known points are necessary to determine it.

A structure with n-point uniqueness cannot have $(n+1)$-point homogeneity. If n points are always sufficient to determine a map, then $n+1$ points cannot be freely assigned. The two properties are closely related, but not identical.

As Narens (1981b) notes, relational structures that can be mapped to an interval scale are 2-point homogeneous and 2-point unique (Theorem 47). When selecting a measurement map, we can begin by mapping any pair of elements to any pair of numbers, and will find the rest of the map uniquely determined from there. Relational structures that correspond to ratio scales, meanwhile, are 1-point homogeneous and 1-point unique under the usual restriction to maps on the positive real numbers, as the selection of any single point determines the scale (Theorem 48). Conversely, and less obviously, *any* relational structure that has 2-point homogeneity and 2-point uniqueness can be represented by an interval scale, and any relational structure that is 1-point homogeneous and 1-point unique can be represented by a ratio scale (although other, non-ratio scale mappings are also possible). The proofs of these last results, in Narens (1981b) and Narens (1981a) respectively, are quite technical and outside of the scope of this paper. But their implications are significant: It means that these scale type properties can usefully group isomorphic mappings. These makes them a valuable classificatory framework for exploring different types of measurement. I will refer to this typology as the "Homogeneity-Uniqueness" (HU) typology.

As a final example of a scale for which the degrees of homogeneity and uniqueness are not equal, we can consider a relational structure similar to that of a ratio scale, but with a

null element included. This element will only have a single possible image in the range set N (called the "true zero" in the case of ratio scales). For example, a bathroom scale may have the option of measuring weight in pounds, kilograms, etc., but will always read 0 when empty. Therefore in this case it is not true that any single element can be mapped to any number (since the null reading can be no other number than 0, and no other weight can be assigned the value of 0). This scale is therefore 0-point homogeneous. This also means that fixing the images of up to 2 points may be necessary to define the whole scale (the null value, and one other value), making the structure 2-point unique. Table 3.1 summarizes these properties.

Scale type	Homogeneity	Uniqueness
Ordinal	∞-point	∞-point
Interval	2-point	2-point
Ratio ($\mathbb{R}^+$)	1-point	1-point
Ratio (incl. 0)	0-point	2-point

Table 3.1: Degrees of homogeneity and uniqueness of common scale types

Using the properties of homogeneity and uniqueness allows us to focus on the process of numerical assignment, rather than the attributes being measured or the resulting numeral series. A relational structure with 2-point homogeneity and 2-point uniqueness is always uniquely determined by arbitrary assignment of any two points to any two numbers. An advantage of the HU typology is that it allows for us to group together different types of measurement maps which are equivalent under these properties.

3.3.2 Extensions to IRT models

3.3.2.1 The scale type of proficiencies

One possible method for assigning proficiency values in a Rasch model was given in Procedure 2. This can be generalized as follows:

Procedure 9.

1. Choose a respondent j and assign a positive real proficiency value t_j.
2. For any other respondent k, the odds ratio of success between k and j will be constant across items (Theorem 29). Call this value c_k.
3. Assign to Respondent k the value $t_k = c_k \cdot t_j$.

This procedure, in which the full scale is determined by the arbitrary assignment of an arbitrary point, exhibits 1-point homogeneity and 1-point uniqueness. It is, in fact, the most

common type of mapping with 1-point homogeneity and 1-point uniqueness: a ratio scale (Theorem 49).

The full Rasch model can be obtained by extending this procedure to assigning difficulty values to items. Given Respondent j with proficiency value t_j, assign item i the following difficulty value:

$$b_i = \frac{t_j}{\text{Odds}(x_i = 1|t_j)} \tag{3.13}$$

This definition is consistent with the Rasch model equation given in Equation 3.2 (Theorem 50). As seen in the previous chapter, by being consistent with the Rasch model, this procedure also constructs a ratio scale, this time on items.

Alternatively, a logit-based procedure can be used to assign proficiency values (Appendix Section C.2.2, Procedure 16). This procedure also involves the full determination of the scale from a single arbitrary assignment, and thus is also 1-point homogeneous and 1-point unique. However, it is an absolute difference scale for proficiencies (Theorem 52). It also establishes an absolute difference scale on items with the following addition:

$$\beta_i = \theta_k - \text{Log odds}(x_i = 1|\theta_k) \tag{3.14}$$

This means that in addition to having the ratio scale type, proficiencies and items in a Rasch model also have the absolute difference scale type, when the logit form of the model equation is used. This example shows one of the advantages of using a homogeneity/uniqueness system. Absolute difference scales and ratio scales are in many ways equivalent to each other, and can be transformed into each other through log or exponential functions. By grouping both in the {1-point homogeneity, 1-point uniqueness} category of the HU typology, that equivalence is made explicit.

The more complex case of 2PL proficiency has two degrees of freedom (2-homogeneity and 2-uniqueness) and can take as its set of measurement maps a variety of scales with these properties. These include an interval scale, with the model given in Equation 3.3. Procedure 7 gave one possible way to assign parameter values. A more generalized version is given by:

Procedure 10.

1. Choose any respondent j_1 and assign a proficiency value θ_{j1}.
2. Choose any respondent j_2 whose observed odds of success on each item are greater than those of j_1, and assign a proficiency value θ_{j2}. (By Theorem 26, if j_2 has greater odds of success than j_1 on any one item, their odds of success will be greater on all items.)

3. For any respondent k, the following ratio will be a constant value for all items (Theorem 36):
$$\frac{(\text{log odds of success for } k) - (\text{log odds of success for } j_1)}{(\text{log odds of success for } j_2) - (\text{log odds of success for } j_1)} \tag{3.15}$$
Call this value c_k.

4. Assign to Respondent k the proficiency
$$\theta_k = \theta_{j1} + c_k(\theta_{j2} - \theta_{j1}) \tag{3.16}$$

To complete the model, for any item i let λ_{j1i} be the observed log odds of success for Respondent j_1, and λ_{j2i} be the observed log odds of success for Respondent j_2. Assign the discrimination parameter
$$\alpha_i = \frac{\lambda_{j2i} - \lambda_{j1i}}{\theta_{j2} - \theta_{j1}} \tag{3.17}$$
and the difficulty parameter
$$\beta_i = \theta_{j1} - \frac{\lambda_{j1i}(\theta_{j2} - \theta_{j1})}{\lambda_{j2i} - \lambda_{j1i}}. \tag{3.18}$$
Then the 2PL model as expressed in Equation 3.3 holds (Theorem 53).

Equivalently, we could use a procedure which is equivalent to the exponentiation of Procedure 10, establishing a relative ratio scale (Appendix, Procedure 17).

3.3.2.2 The scale type of items

Strictly speaking, Procedure 9, even with the addition of Equation 3.13, does not establish the homogeneity or uniqueness of items under a Rasch model, since the procedure begins with the assignment of proficiency values, not item values. An item-based procedure is given by:

Procedure 11.

1. Choose an item i and assign a positive real difficulty value b_i.
2. For any other item l, the odds ratio of success between l and i will be constant across respondents (Theorem 51). Call this value c_l.
3. Assign to Item l the value $b_l = \frac{b_i}{c_l}$.

In this case, the full model can be established by assigning proficiency parameters through this equation:

$$t_j = b_i \cdot \text{Odds}(x_i = 1|t_j) \tag{3.19}$$

Since Procedure 11 assigns the entire map with the arbitrary assignment of a single point, it too has 1-point homogeneity and 1-point uniqueness.

Alternately, item difficulties can be assigned to follow the log odds form of the model (Procedure 18, Appendix Section C.2.3). This procedure is 1-point homogeneous and 1-point unique, and establishes an absolute difference scale.

In the Rasch model case, where items and persons are symmetric with respect to the model, the fact that the two parameter sets have the same scale type is to be expected. In the 2PL model case, the process is more complicated.

If we consider the difficulty and discrimination parameters completely separately, we can use procedures that establish the difficulty side as having the scale type {2-homogenous, 2-unique}, and the discrimination side as having the scale type {1-homogenous, 1-unique}. There are two issues with this approach. One is that the two parameter sets are not fully separable. Choosing the discrimination parameters limits the options for difficulty parameter sets, and vice versa.

The second issue is more subtle. One of the benefits of using this scale typology is that scale type of an attribute is not dependent on a particular parameterization, but relates to properties of the attribute itself. In the Rasch model, for example, the different parameterizations of logit versus odds are considered different scale types in the EDI or SZ typologies, but not in the HU typology. In the 2PL model, the logit v. odds distinction still applies, but even more divergent parameterizations are imaginable.

As seen in Section 3.2.2.1, respondents in a Rasch model can be thought of as magnitudes. As discussed in Section 3.2.3.1, respondents in a 2PL model can be thought of as points on a line. Items in a 2PL model, however, can be thought of as lines in a plane, where the x-axis represents proficiency and the y-axis log odds of success. The usual parameterization of the 2PL from Equation 3.3 (reproduced below) is equivalent to parameterizing a line by giving its slope and x-intercept, since its difficulty parameter represents the proficiency for which log odds of success is 0:

$$\text{logit}(x_i = 1|\theta) = \alpha_i(\theta - \beta_i) \tag{3.3}$$

Alternately, the 2PL could be parameterized with slope and y-intercept:

$$\text{logit}(x_i = 1|\theta) = \alpha_i \cdot \theta + \gamma_i \tag{3.20}$$

where γ_i represents the log odds of success on Item i for a respondent with proficiency parameter $\theta = 0$. Whereas in the model given in Equation 3.3 items were ordered by the amount of proficiency required to have a 50% chance of success, in Equation 3.20 the order is induced by the relative chances of success for a certain respondent (e.g., the mean respondent, if the person scale is mean-centered, but this choice is not obligatory).

The scale type of the γ parameters in Equation 3.20 is very different from that of the β parameters in Equation 3.3. Whereas all possible choices of β parameter maps could be related by linear transformations, the transformations between different γ parameter sets do not even maintain order (Figure 3.3). In a 2PL model, different respondents may order items differently in terms of odds of success. This means that the order of the γ parameters may change depending on which level of proficiency is assigned the proficiency parameter $\theta = 0$.

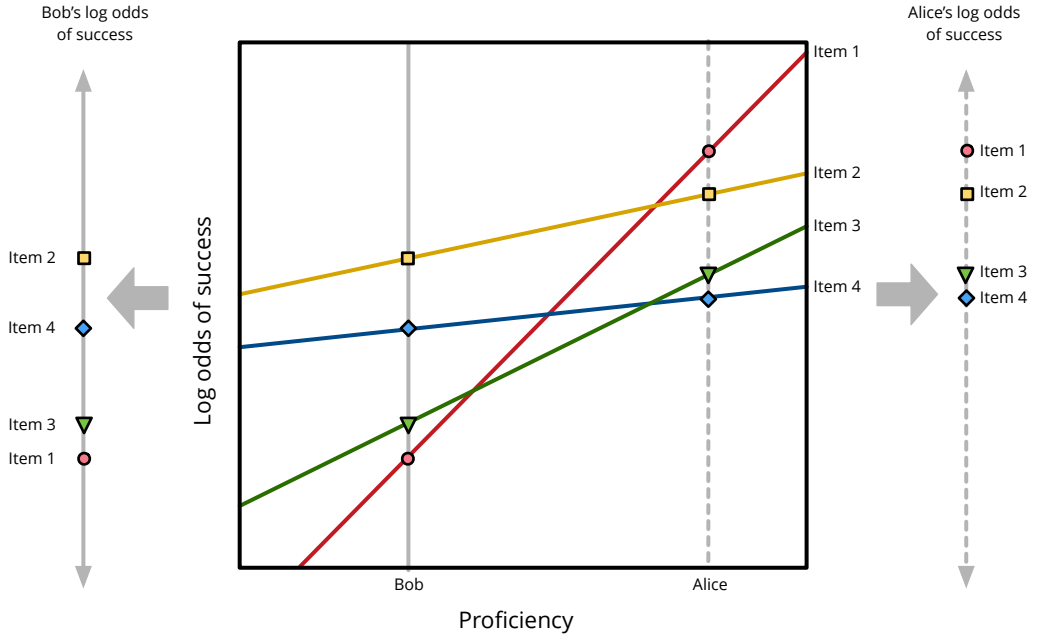


Figure 3.3: Items in a 2PL model have different orders for different respondents. If Alice is assigned the proficiency parameter $\theta_A = 0$, the order of the items under the parameterization in Equation 3.20 will be different from what will result if Bob is assigned the proficiency parameter $\theta_B = 0$.

A scale in which order is not maintained is not interval, or even ordinal. It cannot be classified under the HU typology, which requires a maintained, empirical total order. Overall, a proper scale type of items in the 2PL model would require a theory appropriate to the binary, linear nature of the elements, including a consideration of how the notion of scale type can be expanded to include elements for which representation requires more than one number.

3.4 Summary

Whereas the previous chapter looked at the properties of different types of numerical assignment through the scale types of Suppes and Zinnes (1963), this chapter shows that more

generalized typologies can be used to categorize quantities and their corresponding appropriate scales. Two common types of quantities, magnitudes and points on a line, can be categorized as types {1-homogenous, 1-unique} and {2-homogenous, 2-unique}. In item response theory, the first of these categories includes proficiencies as modeled by the Rasch model, and the second proficiencies as modeled by the 2PL. These categories apply regardless of the form of the IRT models used (odds or logit), and quantities of these types can have a number of different scale types applied, including well known scales such as interval or ratio, or more exotic scales including absolute difference or relative ratio. This can help provide a better way to understand the psychometric quantities being measured.

Appendix A

Chapter 1 proofs and notes

A.1 Difference relations

(from Section 1.3)

Theorem 1. *For any relation which is isomorphic to equality of subtraction under one measurement map (f), there exists another measurement map (g) under which the relation is isomorphic to equality of division. The converse is also true.*

Proof. Following Suppes and Zinnes (1963), let $\mathfrak{U}$ be a relational system with an empirical set X and an empirical equal difference relation. Let $f : \mathfrak{U} \to \mathfrak{R}$, with $\mathfrak{R}$ a real-valued relational system, be a measurement map under which the empirical equal difference relation in $\mathfrak{U}$ is isomorphic to equality of subtraction on $\mathfrak{R}$. Let $a, b, a', b' \in X$ exhibit the empirical equal difference relation. By definition of f, we have $f(a) - f(b) = f(a') - f(b')$. Then we can define $g : X \to \mathbb{R}^+$ such that for all $a \in X$, $g(a) = \exp(f(a))$, and thus $\log(g(a)) = f(a)$. Hence:

$$f(a) - f(b) = f(a') - f(b') \tag{A.1a}$$
$$\log(g(a)) - \log(g(b)) = \log(g(a')) - \log(g(b')) \tag{A.1b}$$
$$\log\left(\frac{g(a)}{g(b)}\right) = \log\left(\frac{g(a')}{g(b')}\right) \tag{A.1c}$$
$$\frac{g(a)}{g(b)} = \frac{g(a')}{g(b')}. \tag{A.1d}$$

□

The converse, that if there exists a measurement map $g : X \to \mathbb{R}^+$ under which a relation is isomorphic to equality of division, it is isomorphic to equality of subtraction under some map $f : X \to \mathbb{R}$, follows analogously if we again let $f(a) = \log(g(a))$ for all $a \in X$.

Theorem 2. *Any property which can be expressed using an absolute ratio scale can be represented on an absolute difference scale in which the assigned numeric values are logs of the ratio scale values.*

Proof. Let the set X with difference relation d be a set representable by an absolute ratio scale where $f : X \to \mathbb{R}$ is the measurement map describing the assignment of real numbers to the elements of X. Then for any $A, B \in X$, the difference factor $d(A, B)$ is real-valued and there exists a continuous bijective function h relating the mathematical ratios $\frac{f(A)}{f(B)}$ to the difference factor $d(A, B)$. This means that just by examining the difference factor $d(A, B)$, it is possible to determine the mathematical difference value $\frac{f(A)}{f(B)}$. In many cases, h will be the identity function, but we will also permit other simple continuous bijective functions.

Let $g : X \to \mathbb{R}$ be the mapping function defined by $g(A) = \log f(A)$ for all $A \in X$. Then for all A, B:

$$h(d(A, B)) = \frac{f(A)}{f(B)} \tag{A.2a}$$

$$\log h(d(A, B)) = \log \frac{f(A)}{f(B)} \tag{A.2b}$$

$$= \log f(A) - \log f(B) \tag{A.2c}$$

$$= g(A) - g(B) \tag{A.2d}$$

$$h'(d(A, B)) = g(A) - g(B) \tag{A.2e}$$

where $h'(x) = \log h(x)$ for all $x \in \mathbb{R}$.

Then there is a continuous bijection between the empirical difference factor $d(A, B)$ and the mathematical difference value $g(A) - g(B)$. If h was the identity, h' is the log function. (If h was exponentiation, h' can be the identity function). This relationship identifies g as an absolute scale. Since the mathematical difference operation is now subtractive, it is an absolute difference scale. □

A.2 IRT relations

(from Section 1.4)

Theorem 3. *Under the odds form of a GRM or 2PL model where item i has discrimination parameter α, the predicted odds ratio on item i between any two respondents A and B with proficiency parameters t_A and t_B respectively will be equal to $\left(\frac{t_A}{t_B}\right)^{\alpha}$.*

Proof.

$$\frac{\text{Odds}(x_i = 1 \mid t_A)}{\text{Odds}(x_i = 1 \mid t_B)} = \frac{\left(\frac{t_A}{b_i}\right)^a}{\left(\frac{t_B}{b_i}\right)^\alpha} \tag{A.3a}$$

$$= \left(\frac{t_A}{t_B}\right)^\alpha \tag{A.3b}$$

□

Corollary.

Theorem 4. *Under a GRM (in which all items have the same discrimination parameter), two respondents A and B will have the same predicted odds ratio on any two items i and j.*

Proof. Follows from Theorem 3. □

Corollary.

Theorem 5. *In the odds form of the Rasch model given in Equation 1.2, the predicted odds ratio of success between two respondents will be equal to the ratio of their proficiency parameters.*

Proof. Follows from Theorem 3 where $\alpha = 1$. □

Corollary.

Theorem 6. *Under the odds form of a GRM or 2PL model, the mathematical ratio $\frac{t_A}{t_B}$ between two respondents' proficiency parameters will be equal to:*

$$\left(\frac{Odds(x_i = 1 \mid t_A)}{Odds(x_i = 1 \mid t_B)}\right)^{\frac{1}{\alpha}} \tag{A.4}$$

where i is any item and α is its discrimination parameter.

Proof. Follows from Theorem 3. □

Theorem 7. *Under the odds form of the 2PL or GRM family of models, if two pairs of respondents have the same odds ratio of success on any item i, then the ratios of their proficiency parameters will also be equal. Conversely, if two pairs of respondents have equal ratios of proficiency parameters, their odds ratios of success on any item will be equal.*

Proof. By Theorem 3:

$$\frac{\text{Odds}(x_i = 1 \mid t_A)}{\text{Odds}(x_i = 1 \mid t_B)} = \frac{\text{Odds}(x_i = 1 \mid t_C)}{\text{Odds}(x_i = 1 \mid t_D)} \tag{A.5a}$$

$$\left(\frac{t_A}{t_B}^{\alpha_i}\right) = \left(\frac{t_C}{t_D}^{\alpha_i}\right) \tag{A.5b}$$

$$\frac{t_A}{t_B} = \frac{t_C}{t_D} \tag{A.5c}$$

The converse follows analogously. □

Corollary.

Theorem 8. *Under the 2PL or GRM family of models, if two pairs of respondents have the same odds ratio on one item, they will have the same odds ratio as each other on all items, even if the items have different discriminations.*

Proof. Follows from Theorem 7. □

Theorem 9. *Under the logit form of a GRM or 2PL model where item i has discrimination parameter α, the predicted odds ratio on item i between any two respondents A and B with proficiency parameters θ_A and θ_B respectively will be equal to $(exp(\theta_A - \theta_B))^{\alpha}$.*

Proof.

$$\frac{\text{Odds}(x_i = 1 \mid \theta_A)}{\text{Odds}(x_i = 1 \mid \theta_B)} = \frac{e^{\text{Log odds}(x_i=1|\theta_A)}}{e^{\text{Log odds}(x_i=1|\theta_B)}} \tag{A.6a}$$

$$= e^{\text{Log odds}(x_i=1|\theta_A) - \text{Log odds}(x_i=1|\theta_B)} \tag{A.6b}$$

$$= e^{\alpha(\theta_A - \beta_i) - \alpha(\theta_B - \beta_i)} \tag{A.6c}$$

$$= e^{\alpha(\theta_A - \theta_B)} \tag{A.6d}$$

$$= \left(e^{\theta_A - \theta_B}\right)^{\alpha} \tag{A.6e}$$

□

Corollary.

Theorem 10. *Under the logit form of a GRM or 2PL model, the mathematical difference $\theta_A - \theta_B$ between two respondents' proficiency parameters will be equal to:*

$$\log\left(\left(\frac{Odds(x_i = 1 \mid \theta_A)}{Odds(x_i = 1 \mid \theta_B)}\right)^{\frac{1}{\alpha}}\right) \tag{A.7}$$

where i is any item and α is its discrimination parameter.

Proof. Follows from Theorem 9. □

Corollary.

Theorem 11. *In the logit form of the Rasch model, the mathematical difference $\theta_A - \theta_B$ between two respondents' proficiency parameters will be equal to the log of their odds ratio of success on any item.*

Proof. Follows from Theorem 10 where $\alpha = 1$. □

Theorem 12. *In the logit form of a GRM or 2PL model, if two pairs of respondents have the same odds ratio of success on any item i, then the subtractive differences of their proficiency parameters will also be equal. Conversely, if two pairs of respondents have equal subtractive differences of proficiency parameters, their odds ratios of success on any item will be equal.*

Proof.

$$\frac{\text{Odds}(x_i = 1 \mid \theta_A)}{\text{Odds}(x_i = 1 \mid \theta_B)} = \frac{\text{Odds}(x_i = 1 \mid \theta_C)}{\text{Odds}(x_i = 1 \mid \theta_D)} \tag{A.8a}$$

$$\log\left(\frac{\text{Odds}(x_i = 1 \mid \theta_A)}{\text{Odds}(x_i = 1 \mid \theta_B)}\right) = \log\left(\frac{\text{Odds}(x_i = 1 \mid \theta_C)}{\text{Odds}(x_i = 1 \mid \theta_D)}\right) \tag{A.8b}$$

$$\begin{aligned}&\log(\text{Odds}(x_i = 1 \mid \theta_A)) - \log(\text{Odds}(x_i = 1 \mid \theta_B)) \\ &= \log(\text{Odds}(x_i = 1 \mid \theta_C)) - \log(\text{Odds}(x_i = 1 \mid \theta_D))\end{aligned} \tag{A.8c}$$

$$\alpha_i(\theta_A - \beta_i) - \alpha_i(\theta_B - \beta_i) = \alpha_i(\theta_C - \beta_i) - \alpha_i(\theta_D - \beta_i) \tag{A.8d}$$

$$\theta_A - \theta_B = \theta_C - \theta_D \tag{A.8e}$$

The converse follows analogously.

□

A.3 Item relations

(from 1.4.3)

Theorem 13. *Under the odds form of the models, if two items with difficulty parameters b_i, b_j have the same discrimination parameter α, the ratio between any participant's predicted odds of success on the two items will be given by:*

$$\left(\frac{b_i}{b_j}\right)^{-\alpha} \tag{A.9}$$

Proof.

$$\frac{\text{Odds}(x_i = 1 \mid t_A)}{\text{Odds}(x_j = 1 \mid t_A)} = \frac{\left(\frac{t_A}{b_i}\right)^{\alpha}}{\left(\frac{t_A}{b_j}\right)^{\alpha}} \tag{A.10a}$$

$$= \left(\frac{b_j}{b_i}\right)^{\alpha} \tag{A.10b}$$

$$= \left(\frac{b_i}{b_j}\right)^{-\alpha} \tag{A.10c}$$

□

Corollary.

Theorem 14. *If two items have the same discrimination parameter, the predicted odds ratio of success between the two items will be independent of the respondent's proficiency level.*

Proof. Follows from Theorem 13. □

Corollary.

Theorem 15. *In the odds form of the GRM, the ratio $\frac{b_i}{b_j}$ between the difficulty parameters b_i, b_j of any two items i, j with common discrimination α will be equal to:*

$$\left(\frac{Odds(x_i = 1 \mid t_A)}{Odds(x_j = 1 \mid t_A)}\right)^{-\frac{1}{\alpha}} \tag{A.11}$$

where A is any respondent and t_A is their proficiency parameter.

Proof. Follows from Theorem 13. □

Corollary.

Theorem 16. *In the odds form of the Rasch model, the ratio $\frac{b_i}{b_j}$ between the difficulty parameters b_i, b_j of any two items i, j will be the reciprocal of the ratio between any respondent's predicted odds of a successful response on the respective items.*

Proof. Follows from Theorem 15 with $\alpha = 1$. □

Theorem 17. *Under the logit form of the models, if two items with difficulty parameters β_i, β_j have the same discrimination parameter α, the subtractive difference between any participant's predicted log odds of success on the two items will be given by:*

$$-\alpha(\beta_i - \beta_j) \tag{A.12}$$

Proof.

$$\text{Log odds}(x_i = 1 \mid \theta_A) - \text{Log odds}(x_j = 1 \mid \theta_A)$$

$$= \alpha(\theta_A - \beta_i) - \alpha(\theta_A - \beta_j) \tag{A.13a}$$
$$= -\alpha(\beta_i - \beta_j) \tag{A.13b}$$

□

Corollary.

Theorem 18. *In the logit form of the GRM, the difference $\beta_i - \beta_j$ between the difficulty parameters β_i, β_j of any two items i, j with common discrimination α will be equal to:*

$$\log\left(\left(\frac{Odds(x_i = 1 \mid \theta_A)}{Odds(x_j = 1 \mid \theta_A)}\right)^{-\frac{1}{\alpha}}\right) \tag{A.14}$$

where A is any respondent and θ_A is their proficiency parameter.

Proof. From Theorem 17:

$$\text{Log odds}(x_i = 1 \mid \theta_A) - \text{Log odds}(x_j = 1 \mid \theta_A) = -\alpha(\beta_i - \beta_j) \tag{A.15a}$$

$$\beta_i - \beta_j = -\frac{1}{\alpha}\left(\text{Log odds}(x_i = 1 \mid \theta_A) - \text{Log odds}(x_j = 1 \mid \theta_A)\right) \tag{A.15b}$$
$$= -\frac{1}{\alpha}\log\frac{\text{Odds}(x_i = 1 \mid \theta_A)}{\text{Odds}(x_j = 1 \mid \theta_A)} \tag{A.15c}$$
$$= \log\left(\left(\frac{\text{Odds}(x_i = 1 \mid \theta_A)}{\text{Odds}(x_j = 1 \mid \theta_A)}\right)^{-\frac{1}{\alpha}}\right) \tag{A.15d}$$

□

Corollary.

Theorem 19. *In the logit form of the Rasch model, the difference $\beta_i - \beta_j$ between the difficulty parameters b_i, b_j of any two items i, j will be the log of the inverse ratio between any respondent's predicted odds of a successful response on the respective items.*

Proof. Follows from Theorem 18 with $\alpha = 1$. □

Theorem 20. *In the logit form of the 2PL model, this value:*

$$\frac{Log\ odds(x_i = 1 \mid \theta_A) - Log\ odds(x_i = 1 \mid \theta_B)}{Log\ odds(x_j = 1 \mid \theta_A) - Log\ odds(x_j = 1 \mid \theta_B)} \tag{A.16}$$

is equal to the ratio of the discrimination parameters for items i and j. This is also true for the odds form of the model (with t_A, t_B substituted for θ_A, θ_B for notational consistency).

Proof. For the log odds form of the model:

$$\frac{\text{Log odds}(x_i = 1 \mid \theta_A) - \text{Log odds}(x_i = 1 \mid \theta_B)}{\text{Log odds}(x_j = 1 \mid \theta_A) - \text{Log odds}(x_j = 1 \mid \theta_B)}$$

$$= \frac{\alpha_i(\theta_A - \beta_i) - \alpha_i(\theta_B - \beta_i)}{\alpha_j(\theta_A - \beta_j) - \alpha_j(\theta_B - \beta_j)} \tag{A.17a}$$

$$= \frac{\alpha_i(\theta_A - \theta_B)}{\alpha_j(\theta_A - \theta_B)} \tag{A.17b}$$

$$= \frac{\alpha_i}{\alpha_j} \tag{A.17c}$$

For the odds form:

$$\frac{\text{Log odds}(x_i = 1 \mid t_A) - \text{Log odds}(x_i = 1 \mid t_B)}{\text{Log odds}(x_j = 1 \mid t_A) - \text{Log odds}(x_j = 1 \mid t_B)}$$

$$= \frac{\log\left(\left(\frac{t_A}{b_i}\right)^{\alpha_i}\right) - \log\left(\left(\frac{t_B}{b_i}\right)^{\alpha_i}\right)}{\log\left(\left(\frac{t_A}{b_j}\right)^{\alpha_j}\right) - \log\left(\left(\frac{t_B}{b_j}\right)^{\alpha_j}\right)} \tag{A.18a}$$

$$= \frac{\alpha_i\left(\log t_A - \log b_i\right) - \alpha_i\left(\log t_B - \log b_i\right)}{\alpha_j\left(\log t_A - \log b_j\right) - \alpha_j\left(\log t_B - \log b_j\right)} \tag{A.18b}$$

$$= \frac{\alpha_i\left(\log t_A - \log t_B\right)}{\alpha_j\left(\log t_A - \log t_B\right)} \tag{A.18c}$$

$$= \frac{\alpha_i}{\alpha_j} \tag{A.18d}$$

□

Corollary.

Theorem 21. *For any two items i, j, this expression:*

$$\frac{\textit{Log odds}(x_i = 1 \mid \theta_A) - \textit{Log odds}(x_i = 1 \mid \theta_B)}{\textit{Log odds}(x_j = 1 \mid \theta_A) - \textit{Log odds}(x_j = 1 \mid \theta_B)} \tag{A.19}$$

is a constant value regardless of the proficiencies of respondents A, B.

Proof. Follows from Theorem 20. □

Appendix B

Chapter 2 proofs and notes

B.1 Scale types

(from Section 2.2.1)

Theorem 22. *Let $f : \mathfrak{U} \to \mathfrak{R}$ be a measurement map for which equal difference relations in $\mathfrak{U}$ are isomorphic to equality of subtraction in $\mathfrak{R}$. For any $g : \mathfrak{U} \to \mathfrak{R}$ such that $g = \phi \circ f$, where $\phi(x)$ is of the form $mx + b$ for real m, b, the isomorphism also applies.*

Proof. Assume that ϕ has the required form. Let $A, B, C, D \in \mathfrak{U}$ have the property that the difference factor between A and B is equal to the difference factor between C and D. Then:

$$
\begin{aligned}
f(A) - f(B) &= f(C) - f(D) && \text{(B.1a)}\\
g(A) - g(B) &= (m \cdot f(A) - b) - (m \cdot f(B) - b) && \text{(B.1b)}\\
&= m \cdot (f(A) - f(B)) && \text{(B.1c)}\\
&= m \cdot (f(C) - f(D)) && \text{(B.1d)}\\
&= (m \cdot f(C) - b) - (m \cdot f(D) - b) && \text{(B.1e)}\\
&= g(C) - g(D) && \text{(B.1f)}
\end{aligned}
$$

Thus, if any two pairs of elements have the equal difference property, and the equality of subtraction isomorphism is true under f, the pairs will have equality of subtraction under g formed by a linear transformation on f. To establish the full isomorphism, let $A, B, C, D \in \mathfrak{U}$ be elements such that $g(A) - g(B) = g(C) - g(D)$. Then:

$$
\begin{aligned}
g(A) - g(B) &= g(C) - g(D) && \text{(B.2a)}\\
(m \cdot f(A) - b) - (m \cdot f(B) - b) &= (m \cdot f(C) - b) - (m \cdot f(D) - b) && \text{(B.2b)}\\
m \cdot (f(A) - f(B)) &= m \cdot (f(C) - f(D)) && \text{(B.2c)}\\
f(A) - f(B) &= f(C) - f(D) && \text{(B.2d)}
\end{aligned}
$$

Since equality of subtraction in g implies equality of subtraction in f, and since equality of subtraction in f implies the equal difference relation in $\mathfrak{U}$, equality of subtraction in g

implies the equal difference relation in $\mathfrak{U}$. Together with the previous result, this means that equality of subtraction in g holds for two pairs of elements if and only if the equal difference relation holds in $\mathfrak{U}$, so the isomorphism holds.

□

Theorem 23. *Let f and g be two measurement maps from $\mathfrak{U}$ to $\mathfrak{R}$ such that for all $A, B \in \mathfrak{U}$, the ratios $\frac{f(A)}{f(B)}$ and $\frac{g(A)}{g(B)}$ are equal. Then there exists real number m such that $g(A) = m \cdot f(A)$ for all $A \in \mathfrak{U}$.*

Proof. For some $A \in \mathfrak{U}$, let $m = \frac{g(A)}{f(A)}$. Then for all $B \in \mathfrak{U}$:

$$\frac{f(A)}{f(B)} = \frac{g(A)}{g(B)} \tag{B.3a}$$

$$\frac{g(B)}{=} \frac{g(A)}{f(A)} \cdot f(B) \tag{B.3b}$$

$$= m \cdot f(B) \tag{B.3c}$$

□

B.2 Double cancellation holds for the Rasch model

Theorem 24. *In a Rasch model, for any $\theta_A, \theta_F, \theta_B, \beta_P, \beta_X, \beta_P$, if $P(x_X = 1|\theta_A) > P(x_Q = 1|\theta_F)$ and $P(x_P = 1|\theta_F) > P(x_X = 1|\theta_B)$ then $P(x_P = 1|\theta_A) > P(x_Q = 1|\theta_B)$.*

Proof.

$$P(x_X = 1 \mid \theta_A) > P(x_Q = 1 \mid \theta_F) \tag{B.4a}$$

$$\text{Logit}(x_X = 1 \mid \theta_A) > \text{Logit}(x_Q = 1 \mid \theta_F) \tag{B.4b}$$

$$\theta_A - \beta_X > \theta_F - \beta_Q \tag{B.4c}$$

$$\theta_A - \theta_F > \beta_X - \beta_Q \tag{B.4d}$$

$$P(\beta_P = 1 \mid \theta_F) > P(\beta_X = 1 \mid \theta_B) \tag{B.4e}$$

$$\theta_F - \beta_P > \theta_B - \beta_X \tag{B.4f}$$

$$\theta_F - \theta_B > \beta_P - \beta_X \tag{B.4g}$$

$$\theta_A - \theta_B > \beta_P - \beta_Q \tag{B.4h}$$

$$\theta_A - \beta_P > \theta_B - \beta_Q \tag{B.4i}$$

$$\text{Logit}(x_P = 1 \mid \theta_A) > \text{Logit}(x_Q = 1 \mid \theta_B) \tag{B.4j}$$

$$P(x_P = 1 \mid \theta_A) > P(x_Q = 1 \mid \theta_B) \tag{B.4k}$$

□

Appendix C

Chapter 3 proofs and notes

C.1 Properties of measure numbers

(from Section 3.2.2.2)

1. The relationship between $[a_1 : b]$ and $[a_2 : b]$ (greater than, equal to, or less than) is the same as the relationship between a_1 and a_2;
2. $[(a + a') : b] = [a : b] + [a' : b]$;
3. $[a : b] = [a : c]/[b : c]$ (and, hence, $[a : b] \cdot [b : c] = [a : c]$;
4. For any unit b and positive real number κ, there exists exactly one magnitude ξ such that $\kappa = [\xi : b]$.
5. $[a : b]$ and $[b : a]$ are multiplicative inverses of each other.
6. $a < b$ if and only if $[a : b] < 1$. $b < a$ if and only if $[a : b] > 1$.

C.2 Additional procedures

C.2.1 Procedures for magnitudes and points on a line

(from Section 3.3.1)

Procedure 12.

1. Choose any point $M \in X$ and any real number to be its image $f(M) \in \mathbb{R}$.
2. Choose any $N \supset M$ and any real number greater than $f(M)$ as its image $f(N) \in \mathbb{R}$.
3. For any point $A \supset M$, define $f(A) = f(M) + (f(N) - f(M)) \times [MA : MN]$.

4. For any point $A \subset M$, define $f(A) = f(M) - (f(N) - f(M)) \times [AM : MN]$.

This definition is consistent for points M and N.

Proof. For $A \neq M, N$, Steps 3 and 4 of the procedure define $f(A) = f(M) \pm (f(N) - f(M)) \times [MA : MN]$. Applying this to M gives:

$$\begin{aligned} f(M) &= f(M) \pm (f(N) - f(M)) \times [MM : MN] && \text{(C.1a)} \\ &= f(M) \pm (f(N) - f(M)) \times 0 && \text{(C.1b)} \\ &= f(M) && \text{(C.1c)} \end{aligned}$$

For N, since by definition $N \supset M$, the appropriate definition is in Step 3: $f(A) = f(M) + (f(N) - f(M)) \times [MA : MN]$. Applying this to N gives:

$$\begin{aligned} f(N) &= f(M) + (f(N) - f(M)) \times [MN : MN] && \text{(C.2a)} \\ &= f(M) + (f(N) - f(M)) \times 1 && \text{(C.2b)} \\ &= f(N) && \text{(C.2c)} \end{aligned}$$

Thus the definition is consistent for both N and M. □

Procedure 13.

1. Choose any point M to map to any positive real number $f(M) \in \mathbb{R}_{>0}$.
2. Choose any point $N \supset M$ to map to any number such that $f(N) > f(M)$.
3. For any $A \supset M$, define $f(A) = f(M) \times \left(\frac{f(N)}{f(M)}\right)^{[MA:MN]}$.
4. For any $A \subset M$, define $f(A) = f(M) \times \left(\frac{f(N)}{f(M)}\right)^{-[AM:MN]}$.

This definition is consistent at M, N.

Proof. For $A \sup M$, Step 3 of the procedure defines $f(A) = f(M) \times \left(\frac{f(N)}{f(M)}\right)^{[MA:MN]}$. Applying this to M gives:

$$\begin{aligned} f(M) &= f(M) \times \left(\frac{f(N)}{f(M)}\right)^{[MM:MN]} && \text{(C.3a)} \\ &= f(M) \times \left(\frac{f(N)}{f(M)}\right)^{0} && \text{(C.3b)} \\ &= f(M) && \text{(C.3c)} \end{aligned}$$

(Using the definition for $A \sup M$ in Step 3 produces the same result.)

For N, since by definition $N \supset M$, the appropriate definition is in Step 3. Applying this to N gives:

$$f(N) = f(M) \times \left(\frac{f(N)}{f(M)}\right)^{[MN:MN]} \tag{C.4a}$$

$$= f(M) \times \left(\frac{f(N)}{f(M)}\right)^{1} \tag{C.4b}$$

$$= f(N) \tag{C.4c}$$

Thus the definition is consistent for both N and M. □

Procedure 14.

1. Choose any quantity b and any positive real number $f(b)$ to be its image under map f.
2. For any other magnitude a, assign $f(b) = [a : b] \cdot f(a)$.

This definition is consistent at b since $[b : b] = 1$.

Procedure 15.

1. Select any quantity b and assign any real number $f(b)$ to be its image.
2. For any other quantity a, assign the value $f(a) = \log([a : b]) + f(b)$.

This procedure is consistent at b since $[b : b] = 1]$.

C.2.2 Procedures for IRT

(from Section 3.3.2)

Procedure 16.

1. Choose a respondent j and assign a positive real proficiency value θ_j.
2. For any other respondent k, the difference in log odds of success between k and j will be constant across items. Call this value d_k.
3. Assign to Respondent k the value $\theta_k = d_k + \theta_j$.

Procedure 17.

1. Choose any respondent j_1 and assign a proficiency value t_{j1}.
2. Choose any respondent j_2 whose observed odds of success on each item are greater than those of j_1, and assign a proficiency value t_{j2}. (By Theorem 26, if j_2 has greater odds of success than j_1 on any one item, their odds of success will be greater on all items.)
3. For any respondent k, the following ratio will be a constant value for all items (Theorem 36):
$$\frac{(\text{log odds of success for } k) - (\text{log odds of success for } j_1)}{(\text{log odds of success for } j_2) - (\text{log odds of success for } j_1)} \tag{C.5}$$
Call this value c_k.
4. Assign to Respondent k the proficiency
$$t_k = t_{j1} \cdot \left(\frac{t_{j2}}{t_{j1}}\right)^{c_k} \tag{C.6}$$

To complete the model, for any item i let λ_{j1i} be the observed log odds of success for Respondent j_1, and λ_{j2i} be the observed log odds of success for Respondent j_2. Assign the discrimination parameter
$$a_i = \frac{\lambda_{j1i} - \lambda_{j21}}{\log\left(\frac{t_{j2}}{t_{j1}}\right)} \tag{C.7}$$
and the difficulty parameter
$$b_i = t_{j1} \cdot \left(\frac{t_{j1}}{t_{j2}}\right)^{\frac{\lambda_{j1i}}{\lambda_{j2i} - \lambda_{ji1}}}. \tag{C.8}$$
This establishes the odds form of the 2PL model.

Proof. For any respondent k and item i, let λ_{ki} be the observed log odds of success for Respondent k. Then

$$t_k = t_{j1} \cdot \left(\frac{t_{j2}}{t_{j1}}\right)^{c_k} \tag{C.9a}$$

$$= t_{j1} \cdot \left(\frac{t_{j2}}{t_{j1}}\right)^{\frac{\lambda_{ki}-\lambda_{j1i}}{\lambda_{j2i}-\lambda_{j1i}}} \tag{C.9b}$$

$$= t_{j1} \cdot \left(\frac{t_{j2}}{t_{j1}}\right)^{-\frac{\lambda_{j1i}}{\lambda_{j2i}-\lambda_{j1i}}} \cdot \left(\frac{t_{j2}}{t_{j1}}\right)^{\frac{\lambda_{ki}}{\lambda_{j2i}-\lambda_{j1i}}} \tag{C.9c}$$

$$= b_i \cdot \left(\frac{t_{j2}}{t_{j1}}\right)^{\frac{\lambda_{ki}}{\lambda_{j2i}-\lambda_{j1i}}} \tag{C.9d}$$

$$\log t_k = \log b_i + \left(\frac{\lambda_{ki}}{\lambda_{j2i} - \lambda_{j1i}}\right) \cdot \log\left(\frac{t_{j2}}{t_{j1}}\right) \tag{C.9e}$$

$$= \log b_i + \lambda_{ki} \cdot \left(\frac{\log\left(\frac{t_{j2}}{t_{j1}}\right)}{\lambda_{j2i} - \lambda_{j1i}}\right) \tag{C.9f}$$

$$= \log b_i + \frac{\lambda_{ki}}{a_i} \tag{C.9g}$$

$$a_i(\log t_k - \log b_i) = \lambda_{ki} \tag{C.9h}$$

$$e^{a_i(\log t_k - \log b_i)} = e^{\lambda_{ki}} \tag{C.9i}$$

$$\left(\frac{e^{\log t_k}}{e^{\log b_i}}\right)^{a_i} = \text{Odds}(x_i = 1 \mid t_k) \tag{C.9j}$$

$$\left(\frac{t_k}{b_i}\right)^{a_i} = \text{Odds}(x_i = 1 \mid t_k) \tag{C.9k}$$

which matches Equation 3.4.

□

C.2.3 Scale types of items

(from Section 3.3.2.2)

Procedure 18.

1. Choose an item i and assign a positive real difficulty value β_i.
2. For any other item l, the difference in log odds of success between i and l will be constant across items. Call this value d_l.
3. Assign to Item l the value $\beta_l = \beta_i - d_l$.

C.3 Theorems

C.3.1 Order independence

(from Section 3.2.2.1, Axiom I)

Theorem 25. *In a Rasch model, if Respondent A's odds of success on item i are higher than Respondent B's odds of success on item i, then for any item j, Respondent A's odds of success will be higher than Respondent B's odds of success.*

Proof.

$$\text{logit}(x_i = 1 \mid \theta_A) > \text{logit}(x_i = 1 \mid \theta_B) \tag{C.10a}$$
$$\theta_A - \beta_i > \theta_B - \beta_i \tag{C.10b}$$
$$\theta_A > \theta_B \tag{C.10c}$$
$$\theta_A - \beta_j > \theta_B - \beta_j \tag{C.10d}$$
$$\text{logit}(x_j = 1 \mid \theta_A) > \text{logit}(x_j = 1 \mid \theta_B) \tag{C.10e}$$

□

Theorem 26. *In a 2PL model, if Respondent A's odds of success on item i are higher than Respondent B's odds of success on item i, then for any item j, Respondent A's odds of success will be higher than Respondent B's odds of success.*

Proof.

$$\text{logit}(x_i = 1 \mid \theta_A) > \text{logit}(x_i = 1 \mid \theta_B) \tag{C.11a}$$
$$\alpha_i(\theta_A - \beta_i) > \alpha_i(\theta_B - \beta_i) \tag{C.11b}$$
$$\theta_A > \theta_B \tag{C.11c}$$
$$\alpha_j(\theta_A - \beta_j) > \alpha_j(\theta_B - \beta_j) \tag{C.11d}$$
$$\text{logit}(x_j = 1 \mid \theta_A) > \text{logit}(x_j = 1 \mid \theta_B) \tag{C.11e}$$

□

C.3.2 Summed odds in a Rasch model

(from Section 3.2.2.1, Axiom III)

Theorem 27. *In a Rasch model, if Respondent A's odds of success on item i are the sum of Respondent B's odds and respondent C's odds on item i, then for any item j Respondent A's odds of success will be the sum of Respondent B's odds and Respondent C's odds on item j.*

Proof.

$$\text{Odds}(x_i = 1 \mid t_A) = \text{Odds}(x_i = 1|t_B) + \text{Odds}(x_i = 1 \mid t_C) \tag{C.12a}$$
$$\frac{t_A}{b_i} = \frac{t_B}{b_i} + \frac{t_C}{b_i} \tag{C.12b}$$
$$t_A = t_B + t_C \tag{C.12c}$$
$$\frac{t_A}{b_j} = \frac{t_B}{b_j} + \frac{t_C}{b_j} \tag{C.12d}$$

□

Theorem 28. *In a Rasch model, if Respondent A's odds of success on item i are the sum of their odds on items j and k, then for any other Respondent B, their odds of success on item i will be the sum of their odds of success on items j and k.*

Proof.

$$\text{Odds}(x_i = 1 \mid t_A) = \text{Odds}(x_j = 1|t_A) + \text{Odds}(x_k = 1 \mid t_A) \tag{C.13a}$$
$$\frac{t_A}{b_i} = \frac{t_A}{b_j} + \frac{t_A}{b_k} \tag{C.13b}$$
$$\frac{1}{b_i} = \frac{1}{b_j} + \frac{1}{b_k} \tag{C.13c}$$
$$\frac{t_B}{b_i} = \frac{t_B}{b_j} + \frac{t_B}{b_k} \tag{C.13d}$$

□

C.3.3 Establishing scales on magnitudes

(from Section 3.2.2.2)

Theorem 29. *In a Rasch model, the predicted odds ratio of success between two respondents is item-independent. In the odds form of the Rasch model given in Equation 3.2, this value will also be equal to the ratio of their proficiency parameters.*

Proof.

$$\frac{\text{Odds}(x_i = 1 \mid t_A)}{\text{Odds}(x_i = 1 \mid t_B)} = \frac{\frac{t_A}{b_i}}{\frac{t_B}{b_i}} \tag{C.14a}$$
$$= \frac{t_A}{t_B} \tag{C.14b}$$
$$= \frac{\frac{t_A}{b_j}}{\frac{t_B}{b_j}} \tag{C.14c}$$
$$= \frac{\text{Odds}(x_j = 1 \mid t_A)}{\text{Odds}(x_j = 1 \mid t_B)} \tag{C.14d}$$

□

Theorem 30. *Assign proficiency values through Procedure 2, starting by assigning $t_B = 1$ to Respondent B. Assign item difficulty values as indicated by Equation 3.9. Then the Rasch model as defined in Equation 3.2 holds.*

Proof. For any respondent A and item i:

$$t_A = \frac{\text{Odds}(x_i = 1 \mid t_A)}{\text{Odds}(x_i = 1 \mid t_B)} \tag{C.15a}$$

$$= \frac{\text{Odds}(x_i = 1 \mid t_A)}{\frac{1}{b_i}} \tag{C.15b}$$

$$\frac{t_A}{b_i} = \text{Odds}(x_i = 1 \mid t_A) \tag{C.15c}$$

which matches the definition in Equation 3.2. □

Theorem 31. *The set of ϕ transformations between any pair of mappings of magnitudes to real numbers that follow Procedure 1 are exactly the set of similarity transformations.*

Proof. Let $\mathfrak{U}$ be the set of quantities together with the order relation and addition operation. The class of allowable functions f that map quantities from $\mathfrak{U}$ to numerals in $\mathfrak{R}$ are distinguished by the choice of quantity b to serve as the unit.

Let b and c be the units associated with the functions f and g respectively. Then for any a, $f(a) = [a : b]$ and $g(a) = [a : c]$. As noted in Section C.1, Hölder shows that for any a, b, c we have $[a : b] = [a : c]/[b : c]$ (Property 3). Then $[a : c] = [a : b] \cdot [b : c]$, and the function ϕ such that $g = \phi \circ f$ is defined by $\phi(x) = x \cdot g$, a similarity transformation.

Conversely, let f be the map defined by setting the unit to be b, and let $\phi(x) = \kappa x$ where κ is some positive real number. Using Property 4, let c be the unique magnitude such that $[c : b] = \kappa$, and thus $\kappa = 1/[b : c]$ (by Property 5). Let $g = \phi \circ f$. Then for any magnitude a, $g(a) = [a : b] \cdot \kappa = [a : b]/[b : c] = [a : c]$, a valid mapping defined by setting the unit to be the magnitude c. □

Theorem 32. *The set of ϕ transformations between any pair of mappings of magnitudes to real numbers that follow Procedure 3 are exactly the set of transformations that add a constant.*

Proof. Let $\mathfrak{U}$ be the set of quantities together with the order relation and addition operation. The class of allowable functions f that map quantities from $\mathfrak{U}$ to numerals in $\mathfrak{R}$ are distinguished by the choice of quantity b to serve as the zero. Let b and c be the zeroes associated with the functions f and g respectively. Then for any a, $f(a) = \log[a : b]$ and $g(a) = \log[a : c]$. As noted in Section C.1, Property 3, Hölder shows that for any a, b, c we have $[a : b] = [a : c]/[b : c]$, so $[a : c] = [a : b] \cdot [b : c]$. Taking the log of both sides gives

us $\log[a:c] = \log[a:b] + \log[b:c]$. Then the function ϕ such that $g = \phi \circ f$ is defined by $\phi(x) = x + f$, the addition of a constant.

Conversely, let f be the map defined by setting the zero to be b, and let $\phi(x) = x + \kappa$ where κ is some real number. Using Property 4, let c be the unique magnitude such that $[c:b] = e^{\kappa}$, and thus $[b:c] = e^{-\kappa}$ by Property 5. This means that $\kappa = -\log[b:c]$. Let $g = \phi \circ f$. Then for any magnitude a:

$$\begin{aligned} g(a) &= \log[a:b] + \kappa && \text{(C.16a)} \\ &= \log[a:b] - \log[b:c] && \text{(C.16b)} \\ &= \log\left(\frac{[a:b]}{[b:c]}\right) && \text{(C.16c)} \\ &= \log[a:c] && \text{(C.16d)} \end{aligned}$$

Therefore g is a valid mapping defined by setting the zero to be the magnitude c. □

C.3.4 Interval relations

(from Section 3.2.3.1)

Equivalence of intervals

Theorem 33. *Under the Rasch model, two respondents A and B will have the same predicted difference in log odds on any item.*

Proof.

$$\begin{aligned} (\theta_A - \beta_i) - (\theta_B - \beta_i) &= \theta_A - \theta_B && \text{(C.17a)} \\ &= (\theta_A - \beta_j) - (\theta_B - \beta_j) && \text{(C.17b)} \\ &= \text{logit}(x_j = 1 \mid \theta_A) - \text{logit}(x_j = 1 \mid \theta_B) && \text{(C.17c)} \end{aligned}$$

□

Theorem 34. *Under the Rasch model, two items i and j will have the same predicted difference in log odds for any respondent.*

Proof.

$$\begin{aligned} (\theta_A - \beta_i) - (\theta_A - \beta_j) &= \beta_j - \beta_i && \text{(C.18a)} \\ &= (\theta_B - \beta_i) - (\theta_B - \beta_j) && \text{(C.18b)} \\ &= \text{logit}(x_i = 1 \mid \theta_B) - \text{logit}(x_j = 1 \mid \theta_B) && \text{(C.18c)} \end{aligned}$$

□

Theorem 35. *Under a 2PL model, the difference in log odds between two respondents A and B is proportional to the item discrimination.*

Proof.

$$\text{logit}(x_i = 1 \mid \theta_A) - \text{logit}(x_i = 1 \mid \theta_B) = (\alpha_i\theta_A - \beta_i) - (\alpha_i\theta_B - \beta_i) \tag{C.19a}$$
$$= \alpha_i(\theta_A - \theta_B) \tag{C.19b}$$

□

Theorem 36. *In a 2PL model, for any four respondents A, B, C, D, the following ratio is a constant value across all items i:*

$$\frac{\textit{logit}(x_i = 1|\theta_A) - \textit{logit}(x_i = 1|\theta_B)}{\textit{logit}(x_i = 1|\theta_c) - \textit{logit}(x_i = 1|\theta_D)} \tag{C.20}$$

Proof.

$$\frac{\text{logit}(x_i = 1 \mid \theta_A) - \text{logit}(x_i = 1 \mid \theta_B)}{\text{logit}(x_i = 1 \mid \theta_c) - \text{logit}(x_i = 1 \mid \theta_D)} = \frac{\alpha_i(\theta_A - \beta_i) - \alpha_i(\theta_B - \beta_i)}{\alpha_i(\theta_C - \beta_i) - \alpha_i(\theta_D - \beta_i)} \tag{C.21a}$$
$$= \frac{\theta_A - \theta_B}{\theta_C - \theta_D} \tag{C.21b}$$
$$= \frac{\alpha_j(\theta_A - \beta_j) - \alpha_j(\theta_B - \beta_j)}{\alpha_j(\theta_C - \beta_j) - \alpha_j(\theta_D - \beta_j)} \tag{C.21c}$$
$$= \frac{\text{logit}(x_j = 1 \mid \theta_A) - \text{logit}(x_j = 1 \mid \theta_B)}{\text{logit}(x_j = 1 \mid \theta_c) - \text{logit}(x_j = 1 \mid \theta_D)} \tag{C.21d}$$

□

Corollary:

Theorem 37. *Under a 2PL model, if two pairs of respondents have the same difference in log odds on one item, they will have the same difference in log odds as each other on all items.*

Proof. Follows from Theorem 36. □

Magnitudes and points on a line

Theorem 38. *Any attribute which conforms to the axioms of magnitude can also be shown to conform to the axioms of points on a line.*

Proof. To apply the axioms of points on a line to a magnitude, it is first necessary to define an order relation $\subset$ and an equivalence class on ordered pairs that will correspond to the notion of the "distance" equivalence classes.

The order relation can be established using the natural definition that $a \subset b$ if and only if $a < b$ (where $<$ is the magnitude order relation). This is enough to fulfill Axiom (α) of the axioms of points on a line (using Axiom II from the axioms of magnitude), as well as

Axiom (β) (using Axiom I), Axiom (γ) (using Hölder's proof of transitivity of $<$ mentioned at the end of Section 3.2.2.1), Axiom δ (using Hölder's proof that $<$ is a dense order), and Axiom κ (using Axiom VII).

The next step is to define the distance equivalency class on intervals, with the interval between two magnitudes a and b represented by the notation ab. The definitions of measure-numbers (Section C.1) will be used to define the distance equivalency as follows: The distances between exhibited by two intervals ab and cd are considered equal if $[a : b] = [c : d]$ or $[a : b] = [d : c]$. Recall that measure-numbers are actual real numbers, so all relations on the real line can be applied.

We will use the following lemmas. For consistency of notation with the axioms of magnitude, $<$ will be used to denote $<$ or $\subset$. The lower case letters from the magnitude definitions will be used to denote our magnitude-location objects.

Lemma 1. *If* $[a : b] = [c : d]$*, then* $[b : a] = [d : c]$*.*

Proof. From Property 5 of Hölder's measure-number properties given in Section C.1, $[a : b] = 1/[b : a]$. Thus:

$$[a : b] = [c : d] \tag{C.22a}$$

$$\frac{1}{[b : a]} = \frac{1}{[d : c]} \tag{C.22b}$$

$$[b : a] = [d : c] \tag{C.22c}$$

□

Lemma 2. *If* $ab = cd$*, then* $cd = ab$*.*

Proof. If $ab = cd$, then $[a : b] = [c : d]$ or $[a : b] = [d : c]$. If $[a : b] = [c : d]$ then $[c : d] = [a : b]$ so $cd = ab$. If $[a : b] = [d : c]$, then by Lemma 1 $[b : a] = [c : d]$ so $[c : d] = [b : a]$ and $cd = ab$. □

Lemma 3. *If* $a < b$*,* $a' < b'$*, and* $ab = a'b'$*, then* $[a : b] = [a' : b']$*.*

Proof. If $ab = a'b'$, then either $[a : b] = [a' : b']$ or $[a : b] = [b' : a']$. By Property 6 of measure-numbers (Section C.1), $[a : b] < 1 < [b' : a']$, so it is not possible that $[a : b] = [b' : a']$. Thus, the only possible equality relation is $[a : b] = [a' : b']$. □

Lemma 4. *If* $ab = cd$*, then* $ab = dc$*.*

Proof. If $ab = cd$, then $[a : b] = [c : d]$ or $[a : b] = [d : c]$. If $[a : b] = [d : c]$ or $[a : b] = [c : d]$, then $ab = dc$. □

Lemma 5. *If* $[a : b] = [a' : b']$ *and* $a < b$*, then* $a' < b'$*.*

Proof. By Property 6 of measure-numbers (Section C.1), $[a : b] < 1$, so $[a' : b'] < 1$, so again by Property 6 $a' < b'$. □

The remaining axioms not discussed in the definition of the order relation are below.

(μ) *Any two intervals can be compared and found to be either equal or unequal.* Follows from Lemma 2.

(ν) *Equality of intervals is transitive.*

Proof. Assume $ab = cd$ and $cd = ef$. Then either $[a : b] = [c : d]$ or $[a : b] = [d : c]$, and either $[c : d] = [e : f]$ or $[c : d] = [f : e]$. Using Lemma 1, the latter means that either $[d : c] = [f : e]$ or $[d : c] = [e : f]$. Using transitivity of equality of real numbers, it then follows that either $[a : b] = [e : f]$ or $[a : b] = [f : e]$, so $ab = ef$, as desired. □

(η) *If $a < b, b < c$ and $a' < b', b' < c'$, then if $ab = a'b'$ and $bc = b'c'$ then $ac = a'c'$.*

Proof. From Lemma 3, $[a : b] = [a' : b']$ and $[b : c] = [b' : c']$. From Property 3 of measure-numbers (Section C.1), $[a : c] = [a : b] \cdot [b : c]$ and $[a' : c'] = [a' : b'] \cdot [b' : c']$. Therefore, $[a : c] = [a' : c']$ and thus $ac = a'c'$. □

(o) If $a < b < c$ and $a' > b' > c'$, then if $ab = a'b'$ and $bc = b'c'$ then $ac = a'c'$.

Proof. By Lemma 4, $ab = b'a'$. By Lemma 3, $[a : b] = [b' : a']$ and $[b : c] = [c' : b']$. Since $[a : c] = [a : b] \cdot [b : c]$ and $[c' : a'] = [c' : b'] \cdot [b' : a']$ (Property 3 of measure-numbers in Section C.1) and multiplication is commutative, $[a : c] = [c' : a']$ and thus $ac = a'c'$. □

(θ) If $m < n$, then for any arbitrary point a, there exists exactly one point b such that $a < b$ and $ab = mn$, and exactly one point c such that $c < a$ and $ca = mn$.

Proof. From Property 4 of measure-numbers (Section C.1), there is exactly one b such that $[b : a] = [n : m]$ (or, by Lemma 1, $[a : b] = [m : n]$), and one c such that $[c : a] = [m : n]$. THen $ab = mn$ and $ca = mn$. Lemma 5, and the fact that $m < n$, together imply that $a < b$ and $c < a$, fulfilling the requirements of the axiom. □

(*) $ab = ba$.

By identity, $[a : b] = [a : b]$, so by definition $ab = ba$.

Thus, all the axioms of points on a line apply to an object that conforms to the axioms of magnitude. □

Theorem 39. *The equivalence class on intervals between points on a line conform to the axioms of magnitude.*

I use "distance" to refer to an equivalence class of intervals. Hölder shows that, provided they are all of the same direction, these distances comply with the axioms of magnitude in Section 3.2.2 with the following definitions for the order relation $<$ and the binary sum function $+$:

Order. Let M, N, M', N' be any points such that $M \subset N$ and $M' \subset N'$. For the two distances a and a' exhibited by intervals MN and $M'N'$ respectively, choose an arbitrary point A. Using axiom (θ), find $B, B' > A$ such that $AB = MN$ and $AB' = M'N'$. This can be imagined as "moving" the intervals on the line to a new common starting point. Then a and a' are ordered based on the order of B and B'. Specifically:

- If $B \subset B'$ then $a < a'$;
- If $B \supset B'$ then $a > a'$;
- If B is the same point as B' then $a = a'$.

Since exactly one of these must be the case, the converse is also true ($a = a'$ implies that B is the same point as B', etc.).

This definition can be extended to define the distances exhibited by intervals of the other direction (NM where $M \subset N$). In this case, after the "moving" operation, define $a < a'$ if $B \supset B'$, and so forth.

Sum. Given interval AB with length x, and MN with length y, the sum $x + y$ can be defined in the following way:

- Choose an arbitrary point P.
- By Axiom (θ), find $Q \supset P$ such that $PQ = AB$ (i.e., PQ has length x.
- Find $R \supset Q$ such that $QR = MN$ (QR has length y).
- By Axiom (o), the length PR is independent of the choice of P (all possible PR formed in this way form a length equivalence class). Then define $z = x + y$ as the length of this interval (and accordingly $z - x = y$).

C.3.5 Defining a scale on locations

(from Section 3.2.3.2)

Theorem 40. *Assign proficiency values through Procedure 7, starting by assigning $\theta_E = 0$ to Respondent E and $\theta_N = 1$ to Respondent N. Assign item discrimination values as indicated by Equation 3.11 and difficulty values as indicated by Equation 3.12. Then the 2PL model as defined in Equation 3.4 holds.*

Proof. For any respondent A and item i:

$$\theta_A = \frac{\text{logit}(x_i = 1 \mid \theta_A) - \text{logit}(x_i = 1 \mid \theta_N)}{\text{logit}(x_i = 1 \mid \theta_E) - \text{logit}(x_i = 1 \mid \theta_N)} \tag{C.23a}$$

$$= \frac{\text{logit}(x_i = 1 \mid \theta_A)}{\text{logit}(x_i = 1 \mid \theta_E) - \text{logit}(x_i = 1 \mid \theta_N)} - \frac{\text{logit}(x_i = 1 \mid \theta_N)}{\text{logit}(x_i = 1 \mid \theta_E) - \text{logit}(x_i = 1 \mid \theta_N)} \tag{C.23b}$$

$$= \frac{\text{logit}(x_i = 1 \mid \theta_A)}{\alpha_i} + \beta_i \tag{C.23c}$$

$$\alpha_i(\theta_A - \beta_i) = \text{logit}(x_i = 1 \mid \theta_A) \tag{C.23d}$$

which matches the definition in Equation 3.4. □

Theorem 41. *For any two maps f and g established through Procedure 6, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then the set of ϕ transformations is exactly the set of linear transformations.*

Proof. Each possible mapping f from locations to numerals is distinguished by its choice of zero-point N and choice of $E \supset N$ to be assigned the numeral 1. For any points N, N', E, E' such that $N \subset E$ and $N' \subset E'$, let f be the mapping such that $f(N) = 0$ and $f(E) = 1$, and let g be the mapping such that $g(N') = 0$ and $g(E') = 1$. Let e be the distance exhibited by NE, let e' be the distance exhibited by $N'E'$, and let n be the distance exhibited by NN'. For any arbitrary point A, let a denote the distance exhibited by NA and AN, and a' the distance exhibited by $N'A$ and AN'. Assume that $N \subset N'$. Then casework can be used to explore each of the possibilities, depending on the placement of A relative to N and N':

1. $N \subset N' \subset A$.

   Since $A \supset N, N'$, $f(A) = [NA : NE]$ and $g(A) = [N'A : N'E']$. Based on this point order, $n + a' = a$ (Theorem 39, Sum), so $a' = a - n$.

$$g(A) = [N'A : N'E'] \tag{C.24a}$$

$$= [a' : e'] \tag{C.24b}$$

$$= [a - n : e'] \tag{C.24c}$$

$$= [a : e'] - [n : e'] \tag{C.24d}$$

$$= [a : e] \cdot [e : e'] - [n : e'] \tag{C.24e}$$

$$= [NA : NE] \cdot [NE : N'E'] - [NN' : N'E'] \tag{C.24f}$$

$$= f(A) \cdot [NE : N'E'] - [NN' : N'E'] \tag{C.24g}$$

2. $N \subset A \subset N'$.

   In this case, $f(A) = [NA : NE]$, and $g(A) = -[AN' : N'E']$. From the point order, $a + a' = n$, so $a' = n - a$. Then:

$$\begin{aligned} g(A) &= -[AN' : N'E'] && \text{(C.25a)} \\ &= -[a' : e'] && \text{(C.25b)} \\ &= -[n - a : e'] && \text{(C.25c)} \\ &= -([n : e'] - [a : e']) && \text{(C.25d)} \\ &= [a : e'] - [n : e'] && \text{(C.25e)} \end{aligned}$$

   From here we can continue the proof of the previous case from Line C.24d.

3. $A \subset N \subset N'$.

   With $A \subset N, N$, $f(A) = -[AN : NE]$ and $g(A) = -[AN' : N'E']$. For the distances, $a + n = a'$. Then:

$$\begin{aligned} g(A) &= -[AN' : N'E'] && \text{(C.26a)} \\ &= -[a' : e'] && \text{(C.26b)} \\ &= -[a + n : e'] && \text{(C.26c)} \\ &= -([a : e'] + [n : e']) && \text{(C.26d)} \\ &= -([a : e] \cdot [e : e'] + [n : e']) && \text{(C.26e)} \\ &= (-[a : e]) \cdot [e : e'] - [n : e'] && \text{(C.26f)} \\ &= (-[AN : NE]) \cdot [NE : N'E'] - [NN' : N'E'] && \text{(C.26g)} \\ &= f(A) \cdot [NE : N'E'] - [NN' : N'E'] && \text{(C.26h)} \end{aligned}$$

   In each case, the function ϕ such that $g = \phi \circ f$ is of the form $\phi = \alpha x + \beta$, where $\alpha = [NE : N'E']$ and $\beta = -[NN' : N'E']$.

   If instead the order is $N' \subset N$, then similar reasoning will yields $g(A) = f(A) \cdot [NE : N'E'] + [N'N : N'E']$. Either way, the function ϕ is a positive linear transformation for any f, g.

   Conversely, let $\phi : \mathbb{R} \to \mathbb{R}$ be a function such that $\phi(x) = \alpha x + \beta$ (where $\alpha > 0$ and $\beta \in \mathbb{R}$). Let N, E be any pair of points and let f be a mapping function as defined in Procedure 6 such that $f(N) = 0$ and $f(E) = 1$. Let e be the distance exhibited by NE and let e' be the distance such that $[e' : e] = \frac{1}{\alpha}$. Let n be the distance such that $[n : e'] = \beta$ (or $-[n : e'] = \beta$ for $\beta < 0$). Then let N' be the point such that $N'N$ has distance n, with $N' \subset N$ for $\beta > 0$ and $N' \supset N$ for $\beta < 0$. Let E' be the point with $E' \supset N'$ such that $N'E'$ has length e'.

Let g be the mapping defined such that $g = \phi \circ f$. Then this gives:

$$g(A) = \phi \circ f(A) \tag{C.27a}$$
$$= \alpha \cdot f(A) + \beta \tag{C.27b}$$
$$= \frac{f(A)}{\frac{1}{\alpha}} + \beta \tag{C.27c}$$

For $\beta < 0$ this gives:

$$g(A) = \frac{f(A)}{[e' : e]} - [n : e'] \tag{C.28a}$$
$$= f(A) \cdot [e : e'] - [n : e'] \tag{C.28b}$$
$$= f(A) \cdot [NE : N'E'] - [NN' : N'E'] \tag{C.28c}$$

From the proofs above (e.g., Lines C.24g and C.26h), this is exactly the expression giving the mapping for an arbitrary point A with a zero-point at $N' \supset N$ and 1 at E'. If $\beta > 0$, then:

$$g(A) = \frac{f(A)}{[e' : e]} + [n : e'] \tag{C.29a}$$
$$= f(A) \cdot [e : e'] + [n : e'] \tag{C.29b}$$
$$= f(A) \cdot [NE : N'E'] + [N'N : N'E'] \tag{C.29c}$$

which as mentioned above is the value at A of a mapping function with zero at $N' \subset N$ and 1 at E'.

Thus, g is a mapping function with $g(N') = 0$ and $g(E') = 1$. Since ϕ was an arbitrary linear transformation, this means that any positive linear transformation on f results in a valid mapping g.

Combined with the above result, this means that the class of linear transformations is exactly the class of possible functions from one map to another.

□

Lemma 6. *For any two points $N \subset E$, let f_{int} be the map defined as in Procedure 6 with $f_{int}(N) = 0$ and $f_{int}(E) = 1$. Let f_{rr} be the map defined as in Procedure 8 with $f_{rr}(N) = 1$ and $f_{rr}(E) = e$. Then for any point A, $f_{rr}(A) = exp(f_{int}(A))$.*

Proof. Per the following table:

Point A	$f_{int}(A)$	$f_{rr}(A)$
N	0	1
E	1	e
$A \supset N$	$[NA:NE]$	$e^{[NA:NE]}$
$A \subset N$	$-[AN:NE]$	$e^{-[AN:NE]}$

For each case, the value of $f_{rr}(A)$ is the exponentiation of $f_{int}(A)$.

□

Theorem 42. *For any two maps f and g established through Procedure 8, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then the set of ϕ transformations is exactly the set of transformations of the form $\phi(x) = \gamma \cdot x^{\alpha}$.*

Proof. For any points N, E, N', E' with $N \subset E$ and $N' \subset E'$, let f_{int} be the map defined using Procedure 6 with $f_{int}(N) = 0$ and $f_{int}(E) = 1$, and g_{int} be the map with $g_{int}(N') = 0$ and $g_{int}(E') = 1$. Let f_{rr} and g_{rr} be the analogous maps using Procedure 8, with $f_{rr}(N) = g_{rr}(N') = 1$, and $f_{rr}(E) = g_{rr}(E') = e$ (the mathematical constant). By Lemma 6, for any point A, $f_{rr}(A)$ and $g_{rr}(A)$ are the exponentiations of $f_{int}(A)$ and $g_{int}(A)$, respectively. From Theorem 41, there exist constant α, β such that $g_{int}(A) = \alpha \cdot f_{int}(A) + \beta$ for all A. Let $\gamma = e^{\beta}$. Then:

$$
\begin{aligned}
g_{rr}(A) &= e^{g_{int}(A)} && \text{(C.30a)}\\
&= e^{\alpha \cdot f_{int}(A) + \beta} && \text{(C.30b)}\\
&= e^{\beta} \cdot \left(e^{f_{int}(A)}\right)^{\alpha} && \text{(C.30c)}\\
&= e^{\beta} \cdot \left(e^{\log f_{rr}(A)}\right)^{\alpha} && \text{(C.30d)}\\
&= e^{\beta} \cdot (f_{rr}(A))^{\alpha} && \text{(C.30e)}\\
&= \gamma \cdot (f_{rr}(A))^{\alpha} && \text{(C.30f)}
\end{aligned}
$$

Conversely, for any α, γ, let $\beta = \log \gamma$ and let ϕ be the function $\phi(x) = \gamma \cdot x^{\alpha}$. For any mapping function f_{rr} defined by Procedure 8 with N, E, such that $f_{rr}(N) = 1$ and $f_{rr}(E) = e$, let g be the function defined by $g = \phi \circ f_{rr}$. Let f_{int} be the mapping function defined by Procedure 6 with $f(N) = 0$ and $f(E) = 1$. By Lemma 6, for any point A, $f_{rr}(A)$ is the exponentiation of $f_{int}(A)$. Then:

$$
\begin{aligned}
g(A) &= \phi \circ f_{rr}(A) && \text{(C.31a)}\\
&= \gamma \cdot (f_{rr}(A))^{\alpha} && \text{(C.31b)}\\
&= e^{\beta} \cdot (f_{rr}(A))^{\alpha} && \text{(C.31c)}\\
&= e^{\beta} \cdot \left(e^{\log f_{rr}(A)}\right)^{\alpha} && \text{(C.31d)}\\
&= e^{\beta} \cdot \left(e^{f_{int}(A)}\right)^{\alpha} && \text{(C.31e)}\\
&= e^{\alpha \cdot f_{int}(A) + \beta} && \text{(C.31f)}
\end{aligned}
$$

By Theorem 41, if $f_{int}(A)$ is a mapping function as defined in Procedure 6, then applying a linear transformation produces another such mapping function. Thus, $g_{int} = \alpha \cdot f_{int}(A) + \beta$ is a mapping function, and Line C.31f can be re-written as $g(A) = e^{g_{int}}$. By Lemma 6, the exponentiation of a mapping function as defined by Procedure 6 is itself a mapping function as defined by Procedure 8. Therefore any ϕ transformation of the form $\phi(x) = \gamma \cdot x^{\alpha}$, when applied to a mapping function following Procedure 6, produces another such mapping function. Combined with the previous result, this means that the set of mapping functions is exactly the set of functions of this type, making this a "relative ratio" scale. □

C.3.6 Generalized procedures

(from Section 3.3.1)

Theorem 43. *For any two maps f and g established through Procedure 12, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then the set of ϕ transformations is exactly the set of linear transformations.*

Proof. Each possible mapping f from locations to numerals is distinguished by its choices of $M, f(M), N$, and $f(N)$. For two such maps f and g, let $M \subset N$ and M', N' be the starting points with images $f(M), f(N), g(M'), g(N')$.

Let n be the distance exhibited by MN, let n' be the distance exhibited by $M'N'$, and let m be the distance exhibited by MM'. For any arbitrary point A, let a denote the distance exhibited by MA and AM, and a' the distance exhibited by $M'A$ and AM'. Assume that $M \subset M'$. Then casework can be used to explore each of the possibilities, depending on the placement of A relative to M and M':

1. $M \subset M' \subset A$.

   Since $A \supset M, M'$, $f(A) = f(M) + (f(N) - f(M)) \cdot [MA : MN]$ and $g(A) = g(M') + (g(N') - g(M')) \cdot [M'A : M'N']$ (Step 3 of Procedure 12). Based on this point order, $m + a' = a$ (Theorem 39, Sum), so $a' = a - m$.

$$g(A) = g(M') + (g(N') - g(M')) \cdot [M'A : M'N'] \tag{C.32a}$$
$$= g(M') + (g(N') - g(M')) \cdot [a' : n'] \tag{C.32b}$$
$$= g(M') + (g(N') - g(M')) \cdot [a - m : n'] \tag{C.32c}$$
$$= g(M') + (g(N') - g(M')) \cdot ([a : n'] - [m : n']) \tag{C.32d}$$
$$= g(M') + (g(N') - g(M')) \cdot ([a : n] \cdot [n : n'] - [m : n']) \tag{C.32e}$$
$$= g(M') + (g(N') - g(M')) \cdot \left(\frac{[n : n']}{f(N) - f(M)} \cdot ((f(N) - f(M)) \cdot [a : n]) - [m : n] \right) \tag{C.32f}$$

$$= g(M') + (g(N') - g(M')) \cdot \left(\frac{[n : n']}{f(N) - f(M)} \cdot (((f(N) - f(M)) \cdot [a : n] + f(M)) - f(M)) - [m : n] \right) \tag{C.32g}$$

$$= g(M') + (g(N') - g(M')) \cdot \left(\frac{[MN : M'N']}{f(N) - f(M)} \cdot (((f(N) - f(M)) \cdot [MA : MN] + f(M)) - f(M)) - [MM' : MN] \right) \tag{C.32h}$$

$$= g(M') + (g(N') - g(M')) \cdot \left(\frac{[MN : M'N']}{f(N) - f(M)} \cdot (f(A) - f(M)) - [MM' : MN] \right) \tag{C.32i}$$

$$= f(A) \cdot \left(\frac{g(N') - g(M')}{f(N) - f(M)} \cdot [MN : M'N'] \right) + \left(g(M') - (g(N') - g(M')) \cdot \left(\frac{f(M) \cdot [MN : M'N']}{f(N) - f(M)} + [MM' : MN] \right) \right) \tag{C.32j}$$

Since $f(M), f(N), g(M'), g(N'), [MN : M'N']$, and $[MM' : MN]$ are all real constants that depend only on the initial choices and not on A, this means that $f(A)$ is shifted by a linear transformation to yield $g(A)$. It remains to be shown that it is the same transformation in all cases.

2. $M \subset A \subset M'$.

   In this case, based on Procedure 12, $f(A) = f(M) + (f(N) - f(M)) \cdot [MA : MN]$ (Step 3), and $g(A) = g(M') - (g(N') - g(M')) \cdot [AM' : M'N']$ (Step 4). From the point order, $a + a' = m$, so $a' = m - a$. Then:

$$\begin{aligned} g(A) &= g(M') - (g(N') - g(M')) \cdot [AM' : M'N'] && \text{(C.33a)} \\ &= g(M') - (g(N') - g(M')) \cdot [a' : n'] && \text{(C.33b)} \\ &= g(M') - (g(N') - g(M')) \cdot [m - a : n'] && \text{(C.33c)} \\ &= g(M') - (g(N') - g(M')) \cdot ([m : n'] - [a : n']) && \text{(C.33d)} \\ &= g(M') + (g(N') - g(M')) \cdot ([a : n'] - [m : n']) && \text{(C.33e)} \end{aligned}$$

   From here we can continue the proof of the previous case from Line C.32d.

3. $A \subset M \subset M'$.

With $A \subset M, M$, $f(A) = f(M) - (f(N) - f(M)) \times [AM : MN]$ and $g(A) = g(M') - (g(N') - g(M')) \cdot [AM' : M'N']$. For the distances, $a + m = a'$. Then:

$$
\begin{aligned}
g(A) &= g(M') - (g(N') - g(M')) \cdot [AM' : M'N'] && \text{(C.34a)}\\
&= g(M') - (g(N') - g(M')) \cdot [a' : n'] && \text{(C.34b)}\\
&= g(M') - (g(N') - g(M')) \cdot [a + m : n'] && \text{(C.34c)}\\
&= g(M') - (g(N') - g(M')) \cdot ([a : n'] + [m : n']) && \text{(C.34d)}\\
&= g(M') - (g(N') - g(M')) \cdot ([a : n] \cdot [n : n'] + [m : n']) && \text{(C.34e)}\\
&= g(M') - (g(N') - g(M')) \cdot \left(\frac{[n : n']}{f(N) - f(M)} \cdot (f(N) - f(M)) \cdot [a : n] + [m : n'] \right) && \text{(C.34f)}\\
&= g(M') - (g(N') - g(M')) \\
&\quad \cdot \left(\frac{[n : n']}{f(N) - f(M)} \cdot (((f(N) - f(M)) \cdot [a : n] - f(M)) + f(M)) + [m : n'] \right) && \text{(C.34g)}\\
&= g(M') - (g(N') - g(M')) \cdot \left(\frac{[MN : M'N']}{f(N) - f(M)} \right. \\
&\qquad \left. \cdot (((f(N) - f(M)) \cdot [MA : MN] - f(M)) + f(M)) + [MM' : M'N'] \right) && \text{(C.34h)}\\
&= g(M') - (g(N') - g(M')) \cdot \left(\frac{[MN : M'N']}{f(N) - f(M)} \cdot (-f(A) + f(M)) + [MM' : M'N'] \right) && \text{(C.34i)}\\
&= f(A) \cdot \left(\frac{g(N') - g(M')}{f(N) - f(M)} \cdot [MN : M'N'] \right) \\
&\quad + \left(g(M') - (g(N') - g(M')) \cdot \left(\frac{f(M) \cdot [MN : M'N']}{f(N) - f(M)} + [MM' : MN] \right) \right) && \text{(C.34j)}
\end{aligned}
$$

This is the same expression as in Line C.32j.

In each case, the function ϕ such that $g = \phi \circ f$ is of the form $\phi = \alpha x + \beta$.

If instead the order is $M' \subset M$, then similar reasoning will yield a slightly modified transformation. Either way, the function ϕ is a positive linear transformation for any f, g.

Conversely, note that in Procedure 12 it is possible to define a mapping function starting with any two points and assigning any two possible values as their images. Then for any real numbers α, β, and any mapping function f defined as in Procedure 12 with starting points M, N and starting values $f(M), f(N)$, define mapping function g with the same starting points M, N and $g(M) = \alpha \cdot (f(M) + \beta, g(N) = \alpha \cdot f(N) + \beta$.

Then for $A \supset M$:

$$g(A) = g(M) + (g(N) - g(M)) \cdot [MA : MN] \tag{C.35a}$$
$$= (\alpha \cdot (f(M) + \beta) + ((\alpha \cdot f(N) + \beta) - (\alpha \cdot (f(M) + \beta)) \cdot [MA : MN] \tag{C.35b}$$
$$= (\alpha \cdot (f(M) + \beta) + (\alpha \cdot f(N) - \alpha \cdot (f(M)) \cdot [MA : MN] \tag{C.35c}$$
$$= \beta + \alpha(f(M) + (f(N) - f(M)) \cdot [MA : MN]) \tag{C.35d}$$
$$= \alpha \cdot f(A) + \beta \tag{C.35e}$$

Thus, g is a mapping function such that $g(A) = \alpha f(A) + \beta$. This means that for any linear transformation ϕ applied to any mapping function f defined through Procedure 12, the result is another valid mapping function. Combined with the above result, this means that the class of linear transformations is exactly the class of possible functions from one map to another.

□

Lemma 7. *For any two points $M \subset N$ and real numbers $r < s$, let f_{int} be the map defined as in Procedure 12 with $f_{int}(M) = r$ and $f_{int}(N) = s$. Let f_{rr} be the map defined as in Procedure 13 with $f_{rr}(M) = e^r$ and $f_{rr}(N) = e^s$. Then for any point A, $f_{rr}(A) = exp(f_{int}(A))$.*

Proof. The lemma holds for M, N by definition. Then for any point $A \supset M$:

$$f_{rr}(A) = f_{rr}(M) \times \left(\frac{f_{rr}(N)}{f_{rr}(M)}\right)^{[MA:MN]} \tag{C.36a}$$
$$= e^r \times \left(\frac{e^s}{e^r}\right)^{[MA:MN]} \tag{C.36b}$$
$$= e^r \times \left(e^{s-r}]\right)^{[MA:MN]} \tag{C.36c}$$
$$= e^r \times e^{(s-r)\cdot[MA:MN]} \tag{C.36d}$$
$$= e^{r+(s-r)\cdot[MA:MN]} \tag{C.36e}$$
$$= e^{f_{int}(M)+(f_{int}(N)-f_{int}(M))\cdot[MA:MN]} \tag{C.36f}$$
$$= e^{f_{int}(A)} \tag{C.36g}$$

For $A \subset M$, the proof is identical, with $-[AM : MN]$ substituted for $[MA : MN]$. □

Theorem 44. *For any two maps f and g established through Procedure 13, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then the set of ϕ transformations is exactly the set of relative ratio transformations $\phi(x) = \gamma \cdot x^\alpha$.*

Proof. For any points $M, N, M'N'$ and real numbers r, s, r', s', with $M \subset N$,$M' \subset N'$,$r < s$ and $r' < s'$, let f_{int} be the map defined using Procedure 12 with $f_{int}(M) = r$ and $f_{int}(N) = s$, and g_{int} be the map with $g_{int}(M') = r'$ and $g_{int}(N') = s$. Let f_{rr} and g_{rr} be the analogous maps using Procedure 13, with $f_{rr}(M) = e^r, f_{rr}(N) = e^s, g_{rr}M' = e^{r'}$, and $g_{rr}N' = e^{s'}$. By Lemma 7, for any point A, $f_{rr}(A)$ and $g_{rr}(A)$ are the exponentiations of $f_{int}(A)$ and $g_{int}(A)$, respectively. From Theorem 43, there exist constant α, β such that $g_{int}(A) = \alpha \cdot f_{int}(A) + \beta$ for all A. Let $\gamma = e^\beta$. Then:

$$\begin{aligned} g_{rr}(A) &= e^{g_{int}(A)} && \text{(C.37a)} \\ &= e^{\alpha \cdot f_{int}(A) + \beta} && \text{(C.37b)} \\ &= e^\beta \cdot \left(e^{f_{int}(A)}\right)^\alpha && \text{(C.37c)} \\ &= e^\beta \cdot \left(e^{\log f_{rr}(A)}\right)^\alpha && \text{(C.37d)} \\ &= e^\beta \cdot (f_{rr}(A))^\alpha && \text{(C.37e)} \\ &= \gamma \cdot (f_{rr}(A))^\alpha && \text{(C.37f)} \end{aligned}$$

Conversely, for any α, γ, let $\beta = \log \gamma$ and let ϕ be the function $\phi(x) = \gamma \cdot x^\alpha$. For any mapping function f_{rr} defined by Procedure 13 with M, N, r, s, such that $f_{rr}(M) = e^r$ and $f_{rr}(N) = e^s$, let g_{rr} be the function defined by $g_{rr} = \phi \circ f_{rr}$. Let f_{int} be the mapping function defined by Procedure 12 with $f(M) = r$ and $f(N) = s$. By Lemma 7, for any point A, $f_{rr}(A)$ is the exponentiation of $f_{int}(A)$. Then:

$$\begin{aligned} g_{rr}(A) &= \phi \circ f_{rr}(A) && \text{(C.38a)} \\ &= \gamma \cdot (f_{rr}(A))^\alpha && \text{(C.38b)} \\ &= e^\beta \cdot (f_{rr}(A))^\alpha && \text{(C.38c)} \\ &= e^\beta \cdot \left(e^{\log f_{rr}(A)}\right)^\alpha && \text{(C.38d)} \\ &= e^\beta \cdot \left(e^{f_{int}(A)}\right)^\alpha && \text{(C.38e)} \\ &= e^{\alpha \cdot f_{int}(A) + \beta} && \text{(C.38f)} \end{aligned}$$

By Theorem 43, if $f_{int}(A)$ is a mapping function as defined in Procedure 12, then applying a linear transformation produces another such mapping function. Thus, $g_{int} = \alpha \cdot f_{int}(A) + \beta$ is a mapping function, and Line C.38f can be re-written as $g_{rr}(A) = e^{g_{int}}$. By Lemma 7, the exponentiation of a mapping function as defined by Procedure 12 is itself a mapping function as defined by Procedure 13. Therefore, any ϕ transformation of the form $\phi(x) = \gamma \cdot x^\alpha$, when applied to a mapping function following Procedure 12, produces another such mapping function. Combined with the previous result, this means that the set of mapping functions is exactly the set of functions of this type, making this a "relative ratio" scale. □

Theorem 45. *For any two maps f and g established through Procedure 14, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then the set of ϕ transformations is exactly the set of similarity transformations.*

Proof. Let f be the map established by defining the image $f(b)$ of b, and g be the map established by defining the image $g(b')$ of b'. Then for any element a:

$$g(a) = [a : b'] \cdot g(b') \tag{C.39a}$$

$$= ([a : b] \cdot [b : b']) \cdot g(b') \tag{C.39b}$$

$$= ([a : b] \cdot f(b)) \cdot [b : b'] \cdot \frac{g(b')}{f(b)} \tag{C.39c}$$

$$= f(a) \cdot [b : b'] \cdot \frac{g(b')}{f(b)} \tag{C.39d}$$

which means that $g(a)$ is equal to $f(a)$ multiplied by a constant (which depends only on the initial choices of $b, b', f(b), g(b')$).

Conversely, let f be the mapping function defined by the image $f(b)$ of b, and let $g(a)$ be equal to $f(a) \cdot c$ for all a (where c is a positive constant). Then:

$$g(b) = f(b) \cdot c \tag{C.40a}$$

$$= [b : b] \cdot f(b) \cdot c \tag{C.40b}$$

$$= f(b) \cdot c \tag{C.40c}$$

and therefore for all a:

$$g(a) = f(a) \cdot c \tag{C.41a}$$

$$= [a : b] \cdot f(b) \cdot c \tag{C.41b}$$

$$= [a : b] \cdot g(b). \tag{C.41c}$$

Thus g can be established through Procedure 14 by defining $g(b) = f(b) \cdot c$. This means that all pairs of mapping functions defined through Procedure 14 are related by a similarity transformation, and all similarity transformations applied to such a mapping results in another such mapping. □

Theorem 46. *For any two maps f and g established through Procedure 15, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then the set of ϕ transformations is exactly the set of transformations that add a constant.*

Proof. Let f be the map established by defining the image $f(b)$ of b, and g be the map established by defining the image $g(b')$ of b'. Then for any element a:

$$\begin{aligned} g(a) &= \log([a \\ &\quad : b']) + g(b') && \text{(C.42a)} \\ &= \log([a \\ &\quad : b] \cdot [b \\ &\quad : b']) + g(b') && \text{(C.42b)} \\ &= \log([a \\ &\quad : b]) + \log([b \\ &\quad : b']) + g(b') && \text{(C.42c)} \\ &= \log([a \\ &\quad : b]) + f(b) - f(b) + \log([b \\ &\quad : b']) + g(b') && \text{(C.42d)} \\ &= f(a) + \log([b \\ &\quad : b']) + g(b') - f(b). && \text{(C.42e)} \end{aligned}$$

So for any pair of maps, the transformation between the two adds a constant. Conversely, let f be the mapping function defined by the image $f(b)$ of b, and let $g(a)$ be equal to $f(a)+c$ for all a (where c is a constant). Then:

$$\begin{aligned} g(b) &= f(b) + c && \text{(C.43a)} \\ &= \log[b \\ &\quad : b] + f(b) + c && \text{(C.43b)} \\ &= f(b) + c && \text{(C.43c)} \end{aligned}$$

and therefore for all a:

$$\begin{aligned} g(a) &= f(a) + c && \text{(C.44a)} \\ &= [a \\ &\quad : b] + f(b) + c && \text{(C.44b)} \\ &= [a \\ &\quad : b] + g(b). && \text{(C.44c)} \end{aligned}$$

Thus g can be established through Procedure 15 by defining $g(b) = f(b) + c$. This means that all pairs of mapping functions defined through Procedure 15 are related by adding a constant, and all such transformations applied to such a mapping results in another such mapping. □

Theorem 47. *Let $\mathscr{X} = \langle X, \succeq, R_1, R_2, \ldots\rangle$ and $\mathscr{N} = \langle N \subseteq \mathbb{R}, S_0, S_1, \ldots\rangle$ be relational structures, and let F be the set of isomorphic measurement maps from $\mathscr{X}$ to $\mathscr{N}$. If this structure defines an interval scale (Suppes & Zinnes, 1963), then $\mathscr{X}$ has 2-homomorphism and 2-uniqueness.*

Proof. Given $f, g \in F$, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then since the structure is an interval scale, the set of ϕ transformations is exactly the set of linear transformations.

Let f be some map in F. To construct another map g, use the following procedure:

Procedure 19.

1. Choose some element $M \in X$ and define its real number image $g(M)$.
2. Choose any point $N \supset M$ and assign any $g(N) > g(M)$.
3. For any $A \in X$, assign

$$g(A) = (f(A) - f(M)) \cdot \frac{g(N) - g(M)}{f(N) - f(M)} + g(M) \tag{C.45}$$

This definition is consistent for M and N:

$$\begin{aligned} g(M) &= (f(M) - f(M)) \cdot \frac{g(N) - g(M)}{f(N) - f(M)} + g(M) && \text{(C.46a)} \\ &= g(M) && \text{(C.46b)} \end{aligned}$$

$$\begin{aligned} g(N) &= (f(N) - f(M)) \cdot \frac{g(N) - g(M)}{f(N) - f(M)} + g(M) && \text{(C.47a)} \\ &= g(N) - g(M) + g(M) && \text{(C.47b)} \\ &= g(N) && \text{(C.47c)} \end{aligned}$$

The g map is then a linear transformation on the f map:

$$\begin{aligned} g(A) &= (f(A) - f(M)) \cdot \frac{g(N) - g(M)}{f(N) - f(M)} + g(M) && \text{(C.48a)} \\ &= f(A) \cdot \left(\frac{g(N) - g(M)}{f(N) - f(M)}\right) + \left(g(M) - f(M) \cdot \frac{g(N) - g(M)}{f(N) - f(M)}\right) && \text{(C.48b)} \end{aligned}$$

This means that $g \in F$ and Procedure 19 has defined a valid map. Additionally, all valid maps in F can be constructed in this way. For any $h \in F$, let $\phi_{hf}(x) = mx + b$. To construct

the h map using Procedure 19, the first two steps assign any $h(M), h(N)$. Then for the third step:

$$h(A) = (f(A) - f(M)) \cdot \frac{h(N) - h(M)}{f(N) - f(M)} + h(M) \tag{C.49a}$$
$$= (f(A) - f(M)) \cdot \frac{(m \cdot f(N) + b) - (m \cdot f(M) + b)}{f(N) - f(M)} + (m \cdot f(M) + b) \tag{C.49b}$$
$$= (f(A) - f(M)) \cdot \frac{(m(f(N) - f(M))}{f(N) - f(M)} + (m \cdot f(M) + b) \tag{C.49c}$$
$$= (f(A) - f(M)) \cdot m + (m \cdot f(M) + b) \tag{C.49d}$$
$$= m \cdot f(A) - m \cdot f(M) + m \cdot f(M) + b \tag{C.49e}$$
$$= m \cdot f(A) + b \tag{C.49f}$$
$$= h(A) \tag{C.49g}$$

So the entire h map is defined in this way. This means that all maps in F can be constructed through Procedure 19, and all maps constructed through Procedure 19 are in F, so the structures are the same. Since the procedure began with selecting two arbitrary points, the underlying structure has 2-homogeneity. Since these steps were enough to define the scale, it has 2-uniqueness. □

Theorem 48. *Let $\mathscr{X} = \langle X, \succeq, R_1, R_2, \ldots \rangle$ and $\mathscr{N} = \langle N \subseteq \mathbb{R}, S_0, S_1, \ldots \rangle$ be relational structures, and let F be the set of isomorphic measurement maps from $\mathscr{X}$ to $\mathscr{N}$. If this structure defines a ratio scale (Suppes & Zinnes, 1963), then $\mathscr{X}$ has 1-homomorphism and 1-uniqueness.*

Proof. Given $f, g \in F$, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then since the structure is an interval scale, the set of ϕ transformations is exactly the set of similarity transformations.

Let f be some map in F. To construct another map g, use the following procedure:

Procedure 20.

1. Choose some element $M \in X$ and define its real number image $g(M)$.
2. For any $A \in X$, assign
$$g(A) = f(A) \cdot \frac{g(M)}{f(M)}$$

This definition is consistent for M:

$$g(M) = f(M) \cdot \frac{g(M)}{f(M)} \tag{C.50a}$$
$$= g(M) \tag{C.50b}$$

The g map is then equal to the f map, multiplied by some constant. This means that $g \in F$ and Procedure 19 has defined a valid map. Additionally, all valid maps in F can be constructed in this way. For any $h \in F$, let $\phi_{hf}(x) = ax$. To construct the h map using Procedure 19, the first step assigns any $h(M)$. Then for the second step:

$$h(A) = f(A) \cdot \frac{h(M)}{f(M)} \tag{C.51a}$$
$$= f(A) \cdot \frac{a \cdot f(M)}{f(M)} \tag{C.51b}$$
$$= f(A) \cdot a \tag{C.51c}$$
$$= h(A) \tag{C.51d}$$

So the entire h map is defined in this way. This means that all maps in F can be constructed through Procedure 20, and all maps constructed through Procedure 20 are in F, so the structures are the same. Since the procedure began with selecting a arbitrary point, the underlying structure has 1-homogeneity. Since this step was enough to define the scale, it has 1-uniqueness.

□

C.3.7 Item response theory

(from Section 3.3.2)

Theorem 49. *For any two maps f and g established through Procedure 9, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then the set of ϕ transformations is exactly the set of similarity transformations, making this a ratio scale.*

Proof. For any two respondents j, j' and real numbers t_j, t'_j, let f and g be the maps defined through Procedure 9 with starting assignments $f(j) = t_j$ and $g(j') = t'_j$. Then for any respondent k, the following holds for any item i:

$$g(k) = \frac{\text{Odds}(x_i = 1 \mid k)}{\text{Odds}(x_i = 1 \mid j')} \cdot g(j') \tag{C.52a}$$
$$g(k) = \frac{\text{Odds}(x_i = 1 \mid k)}{\text{Odds}(x_i = 1 \mid j)} \cdot \frac{\text{Odds}(x_i = 1 \mid j)}{\text{Odds}(x_i = 1 \mid j')} \cdot g(j') \tag{C.52b}$$
$$g(k) = \frac{\text{Odds}(x_i = 1 \mid k)}{\text{Odds}(x_i = 1 \mid j)} \cdot \frac{\text{Odds}(x_i = 1 \mid j)}{\text{Odds}(x_i = 1 \mid j') \cdot f(j)} \cdot \frac{g(j')}{f(j)} \tag{C.52c}$$
$$g(k) = f(k) \cdot \frac{\text{Odds}(x_i = 1 \mid j)}{\text{Odds}(x_i = 1 \mid j')} \cdot \frac{g(j')}{f(j)} \tag{C.52d}$$

By Theorem 29, $\frac{\text{Odds}(x_i=1|j)}{\text{Odds}(x_i=1|j')}$ is a constant that does not depend on the choice of item i. Then for any two functions f and g, g is a transformation of f achieved by multiplying by a constant.

Conversely, note that in Procedure 9 it is possible to define a mapping function starting with any respondent and assigning any possible number as their proficiency value. Then for any real number α, and any mapping function f defined as in Procedure 9 with starting respondent j and assigned proficiency value $f(j)$, define mapping function g with the same starting respondent j and $g(j) = \alpha \cdot f(j)$. Then for any respondent k:

$$g(k) = \frac{\text{Odds}(x_i = 1 \mid k)}{\text{Odds}(x_i = 1 \mid j)} \cdot g(j) \tag{C.53a}$$

$$= \frac{\text{Odds}(x_i = 1 \mid k)}{\text{Odds}(x_i = 1 \mid j)} \alpha \cdot f(j) \tag{C.53b}$$

$$= \alpha \cdot f(k) \tag{C.53c}$$

Thus, given any function f and transformation $\phi(x) = \alpha \cdot x$, the function $g = \phi \circ f$ is also a mapping function. Combined with the above result, this means that all functions of the form $\phi(x) = \alpha \cdot x$ are valid transformations and all valid transformations are of this form. □

Theorem 50. *Assign proficiency values through Procedure 9, starting by assigning t_j to Respondent j. Assign item difficulty values as indicated by Equation 3.13. Then the Rasch model as defined in Equation 3.2 holds.*

Proof. For any respondent k and item i:

$$t_k = \frac{\text{Odds}(x_i = 1 \mid k)}{\text{Odds}(x_i = 1 \mid j)} \cdot t_j \tag{C.54a}$$

$$= \text{Odds}(x_i = 1 \mid k) \cdot \frac{t_j}{\text{Odds}(x_i = 1 \mid j)} \tag{C.54b}$$

$$= \text{Odds}(x_i = 1 \mid k) \cdot b_i \tag{C.54c}$$

$$\frac{t_k}{b_i} = \text{Odds}(x_i = 1 \mid k) \tag{C.54d}$$

which matches the definition in Equation 3.2. □

Theorem 51. *In a Rasch model, the predicted odds ratio of success between two items is person-independent.*

Proof.

$$\frac{\text{Odds}(x_i = 1 \mid t_A)}{\text{Odds}(x_l = 1 \mid t_A)} = \frac{\frac{t_A}{b_i}}{\frac{t_A}{b_l}} \tag{C.55a}$$

$$= \frac{b_l}{b_i} \tag{C.55b}$$

$$= \frac{\frac{t_B}{b_i}}{\frac{t_B}{b_l}} \tag{C.55c}$$

$$= \frac{\text{Odds}(x_i = 1 \mid t_B)}{\text{Odds}(x_l = 1 \mid t_B)} \tag{C.55d}$$

□

Theorem 52. *For any two maps f and g established through Procedure 16, let ϕ_{fg} be the transformation such that $g = \phi_{fg} \circ f$. Then the set of ϕ transformations is the set of functions of the form $\phi(x) = x + c$, making this an absolute difference scale.*

Proof. For any two respondents j, j' and real numbers t_j, t'_j, let f and g be the maps defined through Procedure 16 with starting assignments $f(j) = \theta_j$ and $g(j') = \theta'_j$. Then for any respondent k, the following holds for any item i:

$$\begin{aligned} g(k) &= \text{logit}(x_i = 1 \mid k) - \text{logit}(x_i = 1 \mid j') + \theta_j && \text{(C.56a)} \\ &= (\text{logit}(x_i = 1 \mid k) - \text{logit}(x_i = 1 \mid j) + \theta_j) + \text{logit}(x_i = 1 \mid j) - \text{logit}(x_i = 1 \mid j') && \text{(C.56b)} \\ &= f(k) + \text{logit}(x_i = 1 \mid j) - \text{logit}(x_i = 1 \mid j') && \text{(C.56c)} \end{aligned}$$

By Theorem 33, the difference in log odds $\text{logit}(x_i = 1 \mid j) - \text{logit}(x_i = 1 \mid j')$ is a constant that is independent of the choice of item i. Then the transformation between any two functions consists of adding a constant.

Conversely, given any mapping function f as defined through Procedure 16 with starting assignment $f(j) = \theta_j$, and any real number α, define g as the mapping function with starting assignment $g(j) = \alpha + \theta_j$. Then for any respondent k:

$$\begin{aligned} g(k) &= \text{logit}(x_i = 1 \mid k) - \text{logit}(x_i = 1 \mid j) + g(j) && \text{(C.57a)} \\ &= \text{logit}(x_i = 1 \mid k) - \text{logit}(x_i = 1 \mid j) + \alpha + \theta_j && \text{(C.57b)} \\ &= f(k) + \alpha && \text{(C.57c)} \end{aligned}$$

Thus, given any function f and transformation $\phi(x) = \alpha + x$, the function $g = \phi \circ f$ is also a mapping function. Combined with the above result, this means that all functions of the form $\phi(x) = \alpha + x$ are valid transformations and all valid transformations are of this form. □

Theorem 53. *Assign proficiency values through Procedure 10, starting by assigning θ_{j1} to Respondent j_1 and θ_{j2} to Respondent j_2. Assign item difficulty values as indicated by Equations 3.17 and 3.18. Then the 2PL model as defined in Equation 3.3 holds.*

Proof. For any respondent k and item i:

$$\begin{aligned} \theta_k &= \theta_{j1} + \frac{\text{logit}(x_i = 1 \mid k) - \text{logit}(x_i = 1 \mid j_1)}{\text{logit}(x_i = 1 \mid j_2) - \text{logit}(x_i = 1 \mid j_1)} \cdot (\theta_{j2} - \theta_{j1}) && \text{(C.58a)} \\ &= \theta_{j1} + \frac{\text{logit}(x_i = 1 \mid k) - \lambda_{j1i}}{\lambda_{j2i} - \lambda_{j1i}} \cdot (\theta_{j2} - \theta_{j1}) && \text{(C.58b)} \\ \frac{(\theta_k - \theta_{j1}) \cdot (\lambda_{j2i} - \lambda_{j1i})}{\theta_{j2} - \theta_{j1}} &= \text{logit}(x_i = 1 \mid k) - \lambda_{j1i} && \text{(C.58c)} \\ \text{logit}(x_i = 1 \mid k) &= \frac{(\theta_k - \theta_{j1}) \cdot (\lambda_{j2i} - \lambda_{j1i})}{\theta_{j2} - \theta_{j1}} + \lambda_{j1i} && \text{(C.58d)} \end{aligned}$$

$$= \frac{\lambda_{j2i} - \lambda_{j1i}}{\theta_{j2} - \theta_{j1}} \cdot (\theta_k - \theta_{j1} + \lambda_{j1i} \cdot \frac{\theta_{j2} - \theta_{j1}}{\lambda_{j2i} - \lambda_{j1i}}) \tag{C.58e}$$

$$= \alpha_i(\theta_k - \beta_i) \tag{C.58f}$$

which matches the definition in Equation 3.2. □

Bibliography

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, *21*(1), 1–23.

Andrich, D. (2004). Controversy and the rasch model: A characteristic of incompatible paradigms? *Medical care*, I7–I16.

Artin, M. (1991). *Algebra*. Prentice Hall.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*(4), 351–383.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Borsboom, D. & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory & Psychology*, *14*(1), 105–120.

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, *42*(4), 631–634.

Campbell, N. R. & Jeffreys, H. (1938). Measurement and its importance for philosophy. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *17*, 121–151.

Castellano, K. E., Duckor, B., Wihardini, D., Telléz, K., & Wilson, M. (2016). Assessing academic language in an elementary mathematics teacher licensure exam. *Teacher Education Quarterly*, *43*(1), 3–27.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20.

Cliff, N. (1989). Ordinal consistency and ordinal true scores. *Psychometrika*, *54*(1), 75–91.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*(3), 186–190.

Davey, T., Oshima, T., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, *20*(4), 405–416.

Dedekind, R. (1901). Continuity and irrational numbers. In *Essays on the theory of numbers*. Chicago: Open Court.

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1–19.

Draney, K. L. (1996). *The polytomous saltus model: A mixture model approach to the diagnosis of developmental differences* (Doctoral dissertation, University of California at Berkeley).

Ferrando, P. J. (2009). Difficulty, discrimination, and information indices in the linear factor analysis model for continuous item responses. *Applied Psychological Measurement*, *33*(1), 9–24.

Feuerstahler, L. & Wilson, M. (2019). Scale alignment in between-item multidimensional Rasch models. *Journal of Educational Measurement*, *56*(2), 280–301.

Fischer, G. H. (1995). Derivations of the Rasch model. In *Rasch models* (pp. 15–38). Springer.

Hermisson, S., Gochyyev, P., & Wilson, M. (2018). Assessing pupils' attitudes towards religious and worldview diversity–development and validation of a nuanced measurement instrument. *British Journal of Religious Education*, 1–17.

Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass [The axioms of quantity and the theory of measurement]. In *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig: Mathematische-Physische Classe* [Reports of the Proceedings of the Royal Saxon Society of Sciences in Leipzig: Mathematical-Physical Division] (Vol. 53, pp. 3–64).

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2*(4), 389–423.

Karabatsos, G. (2018). On Bayesian testing of additive conjoint measurement axioms using synthetic likelihood. *Psychometrika*, *83*(2), 321–332.

Keats, J. (1967). Test theory. *Annual Review of Psychology*, *18*(1), 217–238.

Kolstad, A. (1996). The response probability convention embedded in reporting prose literacy levels from the 1992 national adult literacy survey.

Krantz, D. H. & Tversky, A. (1971). Conjoint-measurement analysis of composition rules in psychology. *Psychological Review*, *78*(2), 151.

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, *18*(1), 89–109.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology*, *64*(3), 478–497.

Lord, F. M. (1953). On the statistical treatment of football numbers.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*(1), 15–29.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological bulletin*, *100*(3), 398.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*(5), 639–667.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory & Psychology*, *14*(1), 121–129.

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, *38*(4), 285–294.

Michell, J. (2008). Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory & Psychology*, *18*(1), 119–124.

Michell, J. (2009). The psychometricians' fallacy: Too clever by half? *British Journal of Mathematical and Statistical Psychology*, *62*(1), 41–55.

Michell, J. & Ernst, C. (1996). The axioms of quantity and the theory of measurement: Translated from Part I of Otto Hölder's German text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology*, *40*(3), 235–252.

Michell, J. & Ernst, C. (1997). The axioms of quantity and the theory of measurement: Translated from Part II of Otto Hölder's German text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology*, *41*(4), 345–356.

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, *54*(8), 1024–1048.

Narens, L. (1981a). A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory and Decision*, *13*(1), 1–70.

Narens, L. (1981b). On the scales of measurement. *Journal of Mathematical Psychology*, *24*(3), 249–275.

Narens, L. & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*(2), 166.

Nickerson, C. A. & McClelland, G. H. (1984). Scaling distortion in numerical conjoint measurement. *Applied Psychological Measurement*, *8*(2), 183–198.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821–846.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255.

Preece, P. (2002). Equal-interval measurement: The foundation of quantitative educational research. *Research Papers in Education Policy and Practice*, *17*(4), 363–372.

Rasch, G. (1966). *An informal report on the present state of a theory of objectivity in comparisons*. Universitetets Statistiske Institut.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press. (Original work published 1960)

Reckase, M. D. & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied psychological measurement*, *15*(4), 361–373.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97.

Sbeglia, G. C. & Nehm, R. H. (2019). Do you see what I-SEA? a Rasch analysis of the psychometric properties of the Inventory of Student Evolution Acceptance. *Science Education*, *103*(2), 287–316.

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models (ADISOP). *Psychometrika*, *64*(3), 295–316.

Schwartz, R. & Ayers, E. (2011). *Delta dimensional alignment: Comparing performances across dimensions of the learning progression for Assessing Data Modeling and Statistical Reasoning*. unpublished.

Stenner, A. J. (1994). Specific objectivity—local and general. *Rasch Measurement Transactions*, *8*(3), 374.

Stenner, A. J. (1996). *Measuring reading comprehension with the Lexile framework*. Paper presented at the 4th North American Conference on Adolescent/Adult Literacy. Washington, DC: ERIC.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.

Stevens, S. S. (1959). Measurement, psychophysics, and utility. In C. W. Churchman (Ed.), *Measurement: Definitions and theories*. New York: Wiley.

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, *161*(3844), 849–856.

Stevens, S. S. (1976). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley.

Suppes, P. (1958). *Measurement, empirical meaningfulness, and three-valued logic* (tech. rep. No. 20). OFFICE OF NAVAL RESEARCH.

Suppes, P. & Zinnes, J. L. (1963). Basic measurement theory. *Handbook of mathematical psychology*, *1*(1-76).

Tversky, A. (1967). A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology*, (4), 1–20.

Velleman, P. F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65–72.

Verhelst, N. D. & Glas, C. A. (1995). The one parameter logistic model. In *Rasch models* (pp. 215–237). Springer.

Vessonen, E. (2018). The complementarity of psychometrics and the representational theory of measurement. *The British Journal for the Philosophy of Science*.

Wood, R. (1978). Fitting the Rasch model—a heady tale. *British Journal of Mathematical and Statistical Psychology*, *31*(1), 27–32.

Wright, B. & Linacre, J. (1987). Dichotomous Rasch model derived from specific objectivity. *Rasch measurement transactions*, *1*(1), 5–6.

Yao, S.-Y., Wilson, M., Henderson, J. B., & Osborne, J. (2015). Investigating the function of content and argumentation items in a science test: A multidimensional approach. *Journal of applied measurement*, *16*(2), 171–192.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*(4), 297–311.

Zand Scholten, A. (2011). *Admissible statistics from a latent variable perspective* (Doctoral dissertation, University of Amsterdam).